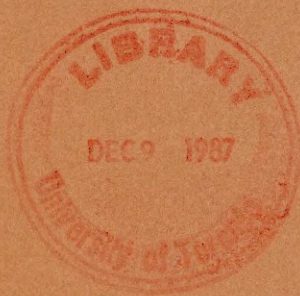


3 1761 10374376 1



SURVEY METHODOLOGY

A JOURNAL
OF
STATISTICS CANADA



VOLUME 13, NUMBER 1
JUNE 1987



Digitized by the Internet Archive
in 2023 with funding from
University of Toronto

<https://archive.org/details/31761103743761>

SURVEY METHODOLOGY

A JOURNAL OF STATISTICS CANADA

JUNE 1987

Published under the authority of
the Minister of Supply and
Services Canada

©Minister of Supply
and Services Canada 1987

Extracts from this publication may be reproduced
for individual use without permission provided the
source is fully acknowledged. However, reproduction
of this publication in whole or in part for purposes
of resale or redistribution requires written permission
from the Publishing Services Group, Permissions
Officer, Canadian Government Publishing Centre,
Ottawa, Canada K1A 0S9

November 1987

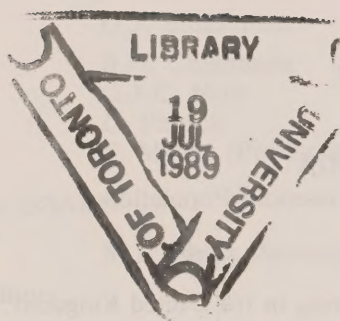
Price: Canada, \$20.00 a year
Other Countries, \$23.00 a year

Payment to be made in Canadian funds or equivalent

Catalogue 12-001, Vol. 13, No. 1

ISSN 0714-0045

Ottawa



Telephone Sample Designs for the U.S. Black Household Population¹

KATHRYN M. INGLIS, ROBERT M. GROVES, and STEVEN G. HEERINGA²

ABSTRACT

The two-stage rejection rule telephone sample design described by Waksberg (1978) is modified to improve the efficiency of telephone surveys of the U.S. Black population. Experimental tests of sample design alternatives demonstrate that: a) use of rough stratification based on telephone exchange names and states; b) use of large cluster definitions (200 and 400 consecutive numbers) at the first stage; and c) rejection rules based on racial status of the household combine to offer improvements in the relative precision of a sample, given fixed resources. Cost and error models are examined to simulate design alternatives.

KEY WORDS: RDD samples; Telephone surveys; Rare population samples.

1. INTRODUCTION

Surveys of rare populations lacking special frames often entail large per-unit costs relative to similar designs for the full population. When the rare population is a small subgroup of a readily identifiable population, the sample of that subgroup is often obtained by screening the larger population. Household surveys of demographic subgroups such as the U.S. Black population typically use such screening to locate eligible sample units; however, extensive screening to identify a rare population sample results in high costs per interview. In recent years telephone-sampling methods have been proposed as cost-efficient tools for sampling and interviewing rare populations. The cost of telephone interviewing is often less than face-to-face interviewing (Groves and Kahn 1979), and when screening is required to identify an eligible respondent, the cost-efficiency of telephone interviewing becomes even more marked. Still, the screening costs of telephone surveys of rare populations can be high in absolute terms.

This paper presents ways in which the screening method for telephone surveys can be refined to reduce costs while achieving desired levels of precision. In this paper we examine a variety of telephone sample designs for the U.S. Black household population. The telephone survey experiments described in this paper were conducted as part of a study of Black political attitudes and electoral behavior in the 1984 U.S. presidential election.

The use of telephone sampling and interviewing implies that Blacks living in households without telephones (about 15 percent of the U.S. Black household population) are not covered by the survey procedures. Such persons tend to be poorer and younger than those living in households with telephones (Thornberry and Massey 1983). To the extent that Blacks without telephones have attitudes and voting behaviors that are different from those with telephones, the survey estimates would differ from Black household population parameters. While not wanting to discount noncoverage error associated with telephone surveys of the Black population, this paper focuses on differential cost efficiencies and sampling error that might result from alternative approaches to telephone samples of Black households.

¹ Revision of paper presented at the 1985 American Statistical Association meetings. Research was partially supported by the U.S. Bureau of the Census and the Survey Research Center. The discussion does not necessarily represent the views of those organizations.

² Kathryn M. Inglis, McNair Anderson and Associates, Australia. Robert M. Groves and Steven G. Heeringa, Survey Research Center, University of Michigan, Ann Arbor, Michigan, 48106-1248, United States.

The telephone sample designs presented here are extensions of a design described by Waksberg (1978). That random digit dialing (RDD) design (commonly referred to as the Waksberg-Mitofsky design) is a two-stage cluster sample of telephone numbers. U.S. telephone numbers contain 10 digits, a three-digit area code, a three-digit central office code or "prefix", and a four-digit suffix in the range 0000-9999 (e.g., 313-764-4424). At the primary stage, a stratified sample of 10-digit telephone numbers is randomly generated, and each such "primary number" is linked to a block of 100 consecutive numbers (e.g., 313-764-4424 would be linked to the "100-series", 313-764-4400 to 313-764-4499). For household surveys, if the primary number is found to be a working household number, then its cluster of 100 consecutive telephone numbers is retained at the first stage for further sampling. If not, its "100-series" is discarded. Therefore, the probability of selection of a first stage 100-series is proportional to the number of working household numbers in that 100-series. In the second stage of sampling, equal numbers of working household numbers are selected from each of the 100-series retained at the primary stage. Therefore, the second stage sampling of households is performed with conditional probabilities of selection inversely proportional to the number of working household numbers in the 100-series. Thus, the design yields an equal probability (epsem) sample of household numbers, and clusters them so that the proportion of total numbers selected which reach households is higher than that obtained by a stratified random RDD sample. To clarify the discussion here, we refer to the 100-series banks of consecutive numbers as the primary stage unit (PSU) of the two-stage RDD design. The term "cluster" is reserved for the fixed set of working household numbers that is selected from the PSUs at the design's second stage.

In this research the sample design modifications aimed at reducing screening costs take three forms: a) stratification of telephone exchange units by proportion Black, and disproportionate allocation of the sample to high density Black strata; b) use of two-stage rejection rules based on both residential status and race of the household; and c) increase in PSU size (from 100 consecutive numbers to 200 and 400).

Stratification of the telephone population by race attempts to isolate exchange areas with high proportions of telephone subscribers who are Black. Higher sampling fractions are then applied to those strata, relative to strata with lower proportions Black. Under this disproportionate sample design, the total number of households that have to be contacted in order to obtain one interview with an eligible Black household is smaller than that for an epsem sample of the household population. Consequently, the screening costs for locating a sample of Black households are reduced. In telephone samples, the basic geographical unit for stratification is the wire center or telephone exchange, to which one or more three-digit prefixes (central office codes) may be assigned. In general, no counts of the subscriber population by racial characteristics are available for these sampling units. Thus, proxy indicators of high density Black exchanges must be used. The experiments described in this paper examined the value of such proxy indicators.

Blair and Czaja (1982) present an alteration of the Waksberg-Mitofsky RDD design which incorporates two-stage rejection rules based on both residential status and race eligibility of the household. For the Black population this method includes, at the first stage, only 100-series whose primary number was assigned to a Black household and then samples a fixed total of Black household numbers within those PSUs. In a U.S. national sample survey, Blair and Czaja found that using this design, the percentage of Black households among all household numbers chosen increased from 9 percent for the first stage to 25 percent for the second stage numbers. Given the compensating probabilities of selection in the two stages, this epsem design greatly reduces the level of screening required to obtain any given sample size of Black households. A similar alteration of the rejection rules for the two-stage Waksberg-Mitofsky design was employed in the experiments described in this paper.

In the Blair and Czaja design some of the primary stage 100-series contained too few Black household numbers to yield the number of elements per cluster required (10 in their case) for an epcem sample of Black households. In addition, relatively large screening costs are incurred at the first stage of selection for this design; over 44 primary numbers must be dialed to locate one Black household. The joint solution to these two problems is to both increase the size of the PSU and to select larger numbers of second stage elements per PSU. The analyses reported here examined the use of primary stage units of 100, 200, and 400 consecutive numbers each. The extension of the PSU definition beyond the standard 100 consecutive numbers was suggested by observations on the assignment of telephone numbers within prefixes. The following appears to be the most common pattern: 1) almost all household numbers within a prefix serve units located within the geographical boundaries of the exchange; 2) there is little geographical clustering of assignments within exchanges (i.e., neighbors do not tend to have consecutive telephone numbers, nor need they have numbers in the same prefix); and 3) there is more diversity in the percentage of household numbers among 1000-series than among 100-series within the same 1000-series of numbers. These impressions are the result of several years of household telephone sampling at the Survey Research Center. Observations 1) to 3) suggest that the expansion of the PSU definition from 100 consecutive numbers to a larger number might permit the use of larger clusters of secondary numbers with little reduction in the proportion of those numbers which are Black households.

2. THE PILOT STUDY

In two integrated experiments imbedded in a pilot survey, several design alternatives were tested. One purpose of the pilot study was to examine the ability of stratification based on civil government units, with only rough correspondence to telephone exchanges, to isolate sets of telephone numbers densely filled with black household numbers. For this, three strata of exchanges were defined:

1. "High density"—Exchanges corresponding to the central cities of large Standard Metropolitan Statistical Areas (e.g., Chicago city, for the Chicago SMSA). This identification was based on the name of the telephone exchanges in these areas.
2. "Medium density"—All other exchanges in selected southern states (Virginia, North Carolina, South Carolina, Florida, Georgia, Alabama, Mississippi, Louisiana). The vast majority of exchanges lie in only one state; those serving two states were associated with the state given in the exchange name.
3. "Low density"—The balance of exchanges in the coterminous United States.

An equal probability sample of 1400 six-digit area code/central office code prefix combinations was then systematically selected from the 34,389 such combinations listed as active on a frame which can be purchased from American Telephone & Telegraph (AT&T). Four-digit random numbers were appended to each selected six-digit stem to yield a sample of 1400 ten-digit primary numbers.

The results of the pilot study demonstrated that the three strata had vastly different proportions of Black telephone numbers. The low density stratum was found to require over six times as much screening to locate a black household as was required in the high density stratum. (This result was confirmed with more precision in the production study, discussed in the next section).

Another purpose of the pilot study was to test the use of rejection rules based on racial composition and working household status of sample numbers from PSUs of differing size. To provide increased precision in analyses related to this objective, an additional 500 primary

numbers were selected from the high- and medium- density strata. The 1900 primary numbers in the combined pilot study sample were then dialed and screened for their Black household status. If the sampled primary number reached a Black household, it simultaneously identified three different PSUs. As shown in Table 1, every individual number can be viewed as belonging to a single 100-series, a single 200-series, and a single 400-series. For example, the number 313-764-4424 is a member of the 4400-4499 100-series, the 4400-4599 200-series, and the 4400-4799 400-series. To test the feasibility of expanding the PSU size, the pilot study sampled secondary numbers from each of these three hundred series. The second stage cluster sizes of Black households were set at 3 for the 100-series of the primary number, 6 for the 200-series, and 9 for the 400-series clusters. In both the primary and secondary stages of selection, if the race of the household was not known, it was assumed to be a non-Black household.

Table 1 presents the disposition of the secondary numbers by PSU type and stratum. Of most interest is the proportion of secondary numbers assigned to Black households for the different PSU definitions. For the 100-series, .134 of all secondary numbers are Black household numbers. This implies that .223 of the households sampled were Black, compared to the .25 Black households found by Blair and Czaja. For the 200-series PSUs, .124 of all secondary numbers are Black household numbers. For the 400-series, .115 of all second stage sample telephone numbers are assigned to Black households. These proportions are all within sampling error of each other (the standard error of each estimate is at least .02). That is, no significant decrease in the proportion eligible was observed when the PSU definition was expanded from 100 to 400 consecutive numbers. These rates imply that while 100-series PSUs on the average can support second stage clusters of 13 or 14 sample Black households, the 400-series might on the average support cluster sizes of 46 sample Black households. The ability to increase the Black household cluster size at the second stage of sampling enables the researcher to greatly reduce sample screening costs.

Table 1 also compares the proportion of eligible secondary numbers for PSUs sampled from the three different strata used in the pilot study. For all the PSU definitions (100, 200, 400) the same result applies — the large SMSA telephone exchanges in the high Black density stratum offer close to a doubling of the eligibility rate when compared to the rate for the overall population (.21 versus .12 or .13). The medium density stratum, consisting of non-SMSA exchanges in selected Southern states, has eligibility rates below that of the nation as a whole (between .08 and .10). The low density stratum, the remainder of the country, also has lower than average eligibility rates (between .07 and .085). Since the high density stratum covers about 36 percent of the Black household population with telephones, the chosen stratification, in combination with disproportionate allocation of the primary stage samples, is an effective tool for reducing screening costs.

3. THE PRODUCTION STUDY

The production study used the stratification plan that was developed and tested in the pilot study. A disproportionately allocated sample of 11,223 primary numbers was selected from the three Black-density strata using sampling fractions in the ratio 3:2:1 (High:Medium:Low). Although the pilot study found no significant difference in the working household rate for PSUs of 200 and 400 consecutive numbers, a conservative decision was made to use the smaller 200-series PSUs in the production study. The expected second stage cluster size for each PSU was set at 5.5 Black households (not counting the primary number). Primary and secondary stage rejection rules for the modified two-stage Waksberg-Mitofsky design were identical to those used for the pilot study. Since much larger sample sizes were used in the production study, questions about precision and relative efficiencies of the design can be addressed with more confidence.

Table 1**Pilot Study**

Disposition of Secondary Numbers Selected within 100-, 200- and 400-Series by Stratum

Stratum and Disposition	Proportion of All Numbers Selected		
	100-Series	200-Series	400-Series*
High Density Black Stratum			
Black Households	.205	.201	.214
Don't Know Race	.028	.029	.032
Non-Black Households	.316	.279	.275
Nonresidential/Nonworking	.451	.491	.479
Number of Cases	(395)	(806)	(1163)
Medium Density Black Stratum			
Black Households	.104	.080	.076
Don't Know Race	.030	.018	.020
Non-Black Households	.494	.443	.420
Nonresidential/Nonworking	.372	.459	.484
Number of Cases	(231)	(560)	(878)
Low Density Black Stratum			
Black Households	.085	.084	.069
Don't Know Race	.014	.028	.027
Non-Black Households	.532	.577	.607
Nonresidential/Nonworking	.369	.311	.297
Number of Cases	(141)	(286)	(491)
Total			
Black Households	.134	.124	.115
Don't Know Race	.024	.025	.026
Non-Black Households	.442	.431	.448
Nonresidential/Nonworking	.400	.420	.411
Number of Cases	(767)	(1652)	(2532)

* Weighted estimate to compensate for the disproportionate allocation of the cluster of 9 secondary numbers across the separate 100-number ranges of the 400-series.

Table 2 presents the results from both the primary and secondary number screening for the production study. The unbiased weighted estimate for an "epsem" two-stage RDD design suggests that 13 percent of all secondary numbers were Black households (the standard error about this estimate is .6 percent). This is in close agreement with the 12 percent secondary number eligibility rate observed in the pilot study. A comparison of the results for the primary stage of selection with those of the secondary stage illustrates the large gains possible by using a two-stage design for telephone sampling of Black households. The gains under the two-stage design are most dramatic in the low density Black stratum where there is nearly a nine-fold increase in the proportion of Black household numbers from the primary to secondary stage (.011 to .090). In the high density stratum the increase is closer to a twofold one (.072 to .190). For the disproportionate allocation design, the unweighted proportions of Black households at the two stages are 3 percent (primary stage) and 15 percent

(secondary stage). Comparison of these figures with the estimates for the epsem design (i.e., 2 percent and 13 percent) indicates the reduction in screening achieved by disproportionate allocation.

As in the pilot study, the percentage of Black households varies over the three strata, although the advantage to distinguishing the medium and low density strata is more evident. Across the three strata, the Black household eligibility rate for secondary numbers varies in an approximate 2:1.5:1 ratio. The three strata also differ in the total proportion of secondary numbers that are assigned to residences. The high density Black stratum has larger proportions of secondary numbers assigned to nonresidential units, probably reflecting the urbanization levels of the exchanges in that stratum.

Table 2
Production Study
Disposition of Numbers Selected by Stratum

Stratum and Disposition	Primaries	Secondaries
High Density Stratum		
Black Households	.072	.190
Don't Know Race	.035	.027
Non-Black Households	.219	.352
Nonresidential/Nonworking	.674	.431
Number of Cases	(3,128)	(6,671)
Medium Density Stratum		
Black Households	.032	.141
Don't Know Race	.020	.018
Non-Black Households	.188	.469
Nonresidential/Nonworking	.760	.372
Number of Cases	(1,879)	(2,375)
Low Density Stratum		
Black Households	.011	.090
Don't Know Race	.019	.023
Non-Black Households	.199	.505
Nonresidential/Nonworking	.771	.382
Number of Cases	(6,116)	(3,987)
Estimate for "Epsem Design"*		
Black Households	.021	.129
Don't Know Race	.021	.023
Non-Black Households	.200	.454
Nonresidential/Nonworking	.758	.394
Proportion Black Households for Disproportionate Design		
	.031	.150
Number of Cases	(11,123)	(13,033)

* Weighted estimates of "epsem design" rates. Weights compensate for disproportionate sampling rates used to select the Production Study sample from the three density strata.

Each PSU of 200 consecutive numbers can be viewed as two half-PSUs of 100 numbers each. Table 3 demonstrates that proportions of nonresidential numbers (.378) found in the half-PSU (100-series) in which the sample primary number fell are lower than in the other half-PSU (.409), but this difference is not statistically significant at the .05 level (standard error about .02). Similarly, the proportion of Black households is somewhat larger in the 100-series of the primary number (.133) than in the adjacent 100-series (.125). Again, this difference is not likely to be found in most replications of the experiment. Table 3 provides another perspective on the results in Table 2, showing only a negligible reduction in the proportion eligible in 100-series adjacent to that of the primary numbers.

The average eligibility rate – proportion of Black households – across PSUs should not be the only criterion for evaluating the sample design. In order to implement an epsem design within strata, each PSU in the design must have a sufficient number of Black households to support the designated number of second stage sample Black households. Thus, the distribution over PSUs of the proportion eligible is also of interest. Figures 1, 2 and 3 contain histograms describing the distribution over all the PSUs of the proportion of Black households by stratum. The stability of the three distributions varies because the number of sample PSUs is about four times greater in the high density stratum than the other two (224 PSUs in the high density stratum to about 60 in the medium and low density strata). The shapes of the distributions, however, appear to be very different for the three strata. The distributions for the low and medium density strata are highly skewed, with 60 percent of PSUs in the medium density stratum and 65 percent of PSUs in the low density stratum having 5 to 20 percent Black households. These eligibility rates correspond to a maximum of 10 to 40 sample Black households for the 200-series PSUs from the low and medium density stratum. In the production study the low density stratum contained several PSUs that would not permit those cluster sizes (6 of the 63 PSUs in that stratum are estimated to have fewer than 10 Black households). The distribution in the high density stratum is much more uniform (4 of the 224 PSUs estimated to have fewer than 10 Black households).

These distributions of percentage Black households by PSU deserve more discussion. Given our current understanding of the assignment of residential numbers to available banks of numbers, there is no reason to believe that within an exchange (or a prefix) there are general tendencies to assign different residential areas to different 100-series. That is, within an exchange serving both Black and non-Black households the hypothesis of assignment of numbers without regard to the race of the subscriber is a strong one. Stated alternatively,

Table 3
Production Study
Disposition of Secondary Numbers by Whether in
Same 100-Series as Primary Numbers

Status	Disposition	
	Same 100-Series as Primary Number	Adjacent 100-Series
Black Households	.133	.125
Don't Know Race	.024	.022
Non-Black Households	.465	.444
Nonresidential/Nonworking	.378	.409
Number of Cases	(6,522)	(6,511)

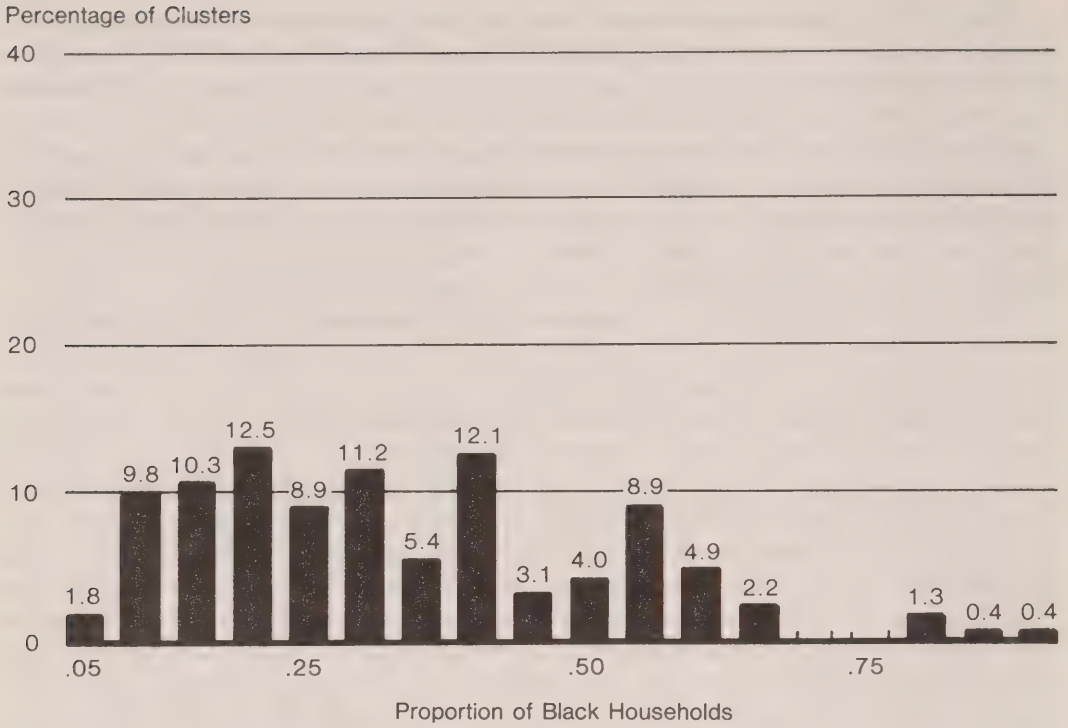


Figure 1. Percentage of High Density Clusters By Proportion of Black Households

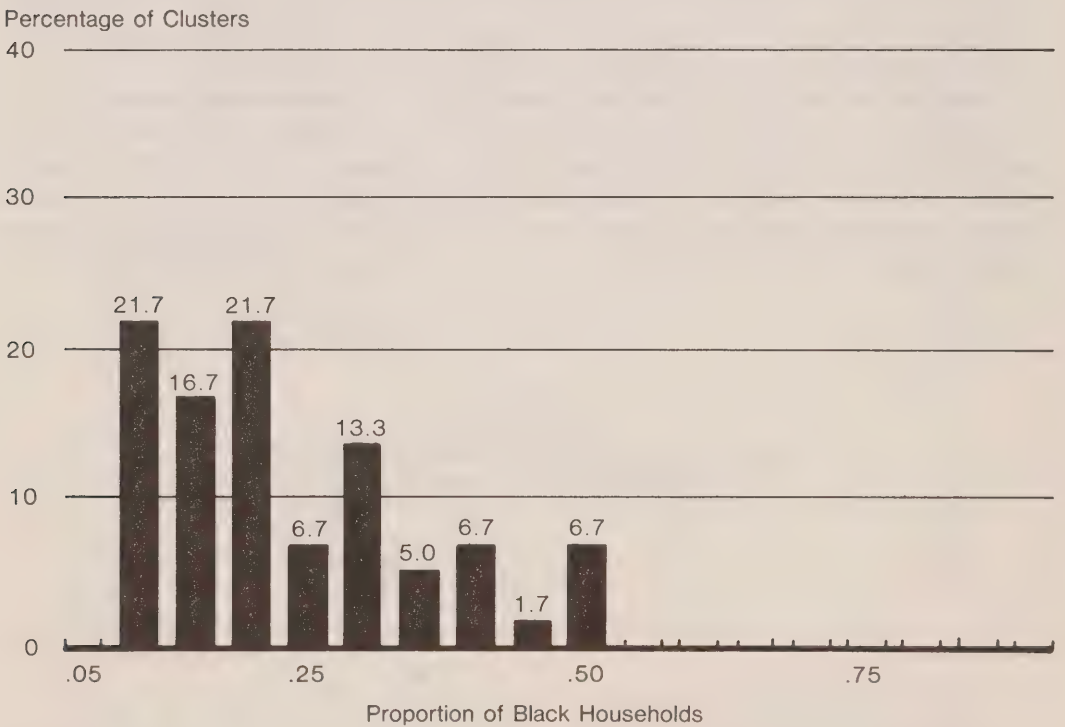


Figure 2. Percentage of Medium Density Clusters By Proportion of Black Households

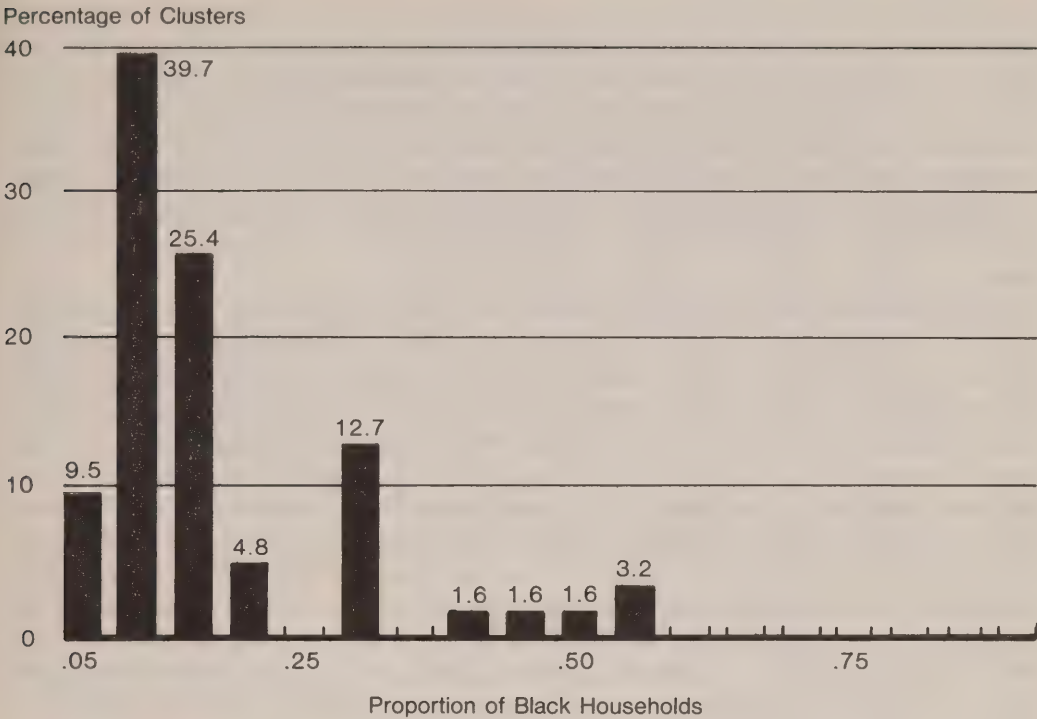


Figure 3. Percentage of Low Density Clusters By Proportion of Black Households

unless the exchanges are subdivided into wire centers that correspond to the residential locations of Black households, there is no *a priori* reason for large amounts of clustering of Black households within 200-series. Following this logic, the more uniform distribution in the high density stratum reflects, we believe, the variability in proportions of Blacks among the telephone populations in the different exchanges in the stratum.

4. SAMPLING VARIANCE PROPERTIES

To achieve greater cost-efficiency in the RDD sampling of Black households it is advantageous to use both large clusters of sample households per PSU (i.e., for a fixed sample size, a smaller number of PSUs) and disproportionate allocation of PSUs to strata of exchanges which vary in their proportion of Black telephone households. While both greater clustering and disproportionate allocation of the sample improve cost-efficiency, the overall precision of the sample is affected by the increased clustering effects and added design effects due to the non-optimal weighting that is required to compensate for the unequal selection probabilities for households from the three density strata. Increased design effects of sample estimates due to non-optimal weighting are described in Kish (1976). The clustering influence on the design effect for the modified RDD procedures is developed in the following paragraphs.

Ceteris paribus, the larger the number of sample elements chosen per PSU the higher the design effect (the ratio of the sampling variance of the given design to that of a simple random sample with the same number of elements). The model often used is $Deff = 1 + \rho(b - 1)$, where *Deff* is the design effect, ρ is the intracluster correlation for

the statistic, and b is the number of sample elements per PSU. Others have shown for many variables on the total U.S. household population that the intracluster correlations for the 100-series tend to be smaller than those generally found in area probability sample clusters (see Groves, 1978). This may not be the case for the Black population for 100-series, and there are no empirical estimates available concerning intracluster correlations for 200-series clusters. The expectation prior to estimating sampling errors was that there would be no change in the intracluster correlations between the 100- and 200-series. This hypothesis reflects the understanding of the assignment of telephone numbers within exchanges that was described above.

Based on sampling errors estimated from the production study data set, the average design effect for a selected set of seven survey statistics is 1.28 for the 100-series and 1.30 for the 200-series. The 100-series average design effect was estimated from those cases which fell into the 100-series of the primary number, while the cases from the entire 200-series were used in computing the average 200-series design effect. Thus, the average cluster size of completed interviews is 2.0 for the 100-series (coefficient of variation, .043) and 3.4 for the 200-series (coefficient of variation, .029). These design effects reflect all the stratification, clustering and weighting in the design and also the fact that the variability in the cluster sizes in the 100-series is greater. (The rejection rule forced an equal number of sample Black households at the 200-series but not necessarily at the 100-series level.) Given that the average design effects for the 100-series and the 200-series are close to one another (1.28 to 1.30), the dominant influence on the sampling variance appears to be non-optimal weighting required by the disproportionately allocated sample design, with little loss in precision due to PSU size alone (moving from the 100- to the 200-series clusters).

Table 4 (page 11) presents the synthetic intracluster correlations by stratum for the seven survey statistics used to compute the estimate of average design effect. The estimates of synthetic intracluster correlations were obtained from the design effect, following Kish's model of $Rho = (Deff - 1)/(b - 1)$, and are unweighted so as to remove the confounding effect of weighting on the synthetic estimates. The estimates in the table tend to be unstable due to the small number of clusters in each stratum, the small average cluster size of completed interviews, and its associated coefficient of variation. These sample design features complicate our inference about clustering effects in the 100- versus the 200-series. Overall, the 100-series estimates of intracluster correlation are somewhat higher than those in the 200-series. We believe that this reflects more an instability in the estimated synthetic correlation than a real difference in clustering effects. We believe that these estimates provide little evidence that there is a change in the intracluster correlation between the 100- and 200-series.

5. OPTIMAL DESIGN FEATURES

The previous sections of the paper address the effect of alternative sample features on cost-efficiency and sampling variance. Survey costs and errors are often combined at the design step to address whether "optimal" features of the survey can be identified. This approach attempts to identify the design which offers minimum variance for a fixed set of resources allocated to the survey. Given the data in this research we can estimate the optimal choices of two design attributes: a) number of sample elements per PSU, and b) allocation of the sample across the three "Black-density" strata.

To determine the optimal cluster size we use a total cost model, $C = C_o + C_a a + C_b ab$, where C_o represents fixed costs, C_a is the sampling and screening cost for each sample cluster, of which a are selected, and C_b is the sampling, screening and interviewing cost

Table 4
Production Study
Synthetic Intracluster Correlations
for 100- and 200-Series Clusters for Seven Statistics by Stratum

Statistic	Synthetic Intracluster Correlation*					
	High Density Black Stratum		Medium Density Black Stratum		Low Density Black Stratum	
	100- Series	200- Series	100- Series	200- Series	100- Series	200- Series
Proportion Very Satisfied with Life as a Whole	.021	-.002	-.172	-.042	-.238	-.116
Proportion Who Think They Are Better Off Financially Than One Year Ago	.113	.075	.094	.069	.206	.049
Proportion Who Will Vote for Mondale	.189	.021	.086	-.087	-.436	-.046
Proportion Who Attend Church	.013	.017	-.009	-.078	.035	-.110
Proportion in Same City or Town All of Life	-.078	.001	.058	.114	.221	.248
Proportion Voted in 1980 Presidential Election	-.045	-.035	-.101	-.013	.364	.356
Proportion Who Think Reagan Will Be Elected President	-.045	-.045	-.545	-.078	.124	-.105
Average	.024	.005	-.084	-.016	.039	.039

* These estimates are unweighted.

associated with each interview obtained, of which there are b in each cluster. Because the proportions of Black households vary across the three strata in the design, the C_a and C_b parameters vary across strata (see Table 5). The optimal cluster size is computed as $\sqrt{C_a(1 - \rho)/(C_b\rho)}$ (Kish, 1965). Using cost data from the production survey, Table 5 presents estimated optimal cluster sizes for overall means and proportions with three alternative levels of intracluster correlation; .005, .01, and .02. (These values are similar to those obtained for attitudinal and behavioral variables in the actual surveys.) The C_a and C_b cost estimates for each stratum also appear. The Table shows that the optimal cluster sizes are largest in the low density stratum, reflecting the high screening costs in that group. Note also that these optimal cluster sizes tend to be larger than those actually used in the survey, $\bar{b} = 6.5$.

Note further that the optimal cluster sizes are similar for 100- and 200-series PSUs and the loss of cost-efficiency of the 200-series relative to that of the 100-series is minor and similar optimal cluster sizes result. (The sampling variance estimates also imply that intracluster correlations in the 100- and 200-series clusters are similar.)

The optimal cluster sizes in Table 5 generally exceed the levels that could be supported with a 100-series PSU definition. That is, a large proportion of 100-series PSUs would not have a sufficient number of Black household numbers to fulfill the designated second stage cluster size. For that reason alone, the 200-series is favored. Even with 200-series, the specified second stage cluster sizes could not be obtained for some PSUs in the low density stratum. (This suggests the true optimal cluster size solution should be constrained to reflect the capacities of the PSUs and the approach used here is useful to guide practical decisions on cost-efficiency, but does not reflect some extreme conditions.)

Table 5
Cost Parameters and Optimal Number of Sample Elements Per Cluster,
by Stratum for 100- and 200-Series Clusters and Different ρ Values

Stratum and Cluster Definition	Optimal Cluster Size			Cost Parameters	
	$\rho = .005$	$\rho = .01$	$\rho = .02$	C_{ha}	C_{hb}
High Density Stratum					
100	15.9	11.2	7.9	\$50.81	\$40.11
200	15.9	11.2	7.9		\$39.78
Medium Density Stratum					
100	22.4	15.8	11.1	\$114.09	\$45.18
200	21.3	15.0	10.6		\$50.00
Low Density Stratum					
100	29.8	21.0	14.8	\$309.98	\$69.52
200	29.9	21.1	14.8		\$69.18

The second design decision evaluated is the choice of sample allocation to strata. The survey used sampling fractions in the ratio of 3:2:1 from the high density to the low density stratum. We explored the optimal allocation across strata, assuming that the optimal cluster sizes were chosen in each stratum (as shown in Table 5). Given a fixed cluster size in each stratum, b_h , we set the sampling fraction in the h -th stratum, f_h , proportional to $\sqrt{(Deff_h S_h^2)/(C_{ha}/b_h)}$, where $Deff_h$ is the design effect for the statistic in the h -th stratum, S_h^2 is the element variance in the h -th stratum, C_{ha} is the sampling and screening costs for PSUs in the h -th stratum, and b_h is the number of sample elements per cluster in the h -th stratum.

Table 6 presents optimal ratios of sampling fractions for various combinations of element variances in the three strata and the various ρ values. The Table shows that the optimal allocations across strata are relatively insensitive to changes in ρ values (for the range of ρ values that are likely given this design). If the strata with higher densities of Black households have element variances at least equal to that of the low density stratum, an oversampling of those strata is desirable. (This reflects the much lower costs in those strata.) The 3:2:1 ratio of sampling fractions is best when the ratio of strata standard deviations is about 1.7:1.5:1. An examination of the data obtained from the survey suggests that many variables have ratios of standard deviations across the three strata close to 1:1:1. For such variables the optimal ratio of sampling fractions is 1.7:1.4:1, given the optimal cluster sizes shown in Table 5. (With the cluster size of 6.5 actually used in each stratum, the optimal fractions have the ratio 2.5:1.6:1.) Both these ratios of sampling fractions suggest that the oversampling actually used in the production study created a loss of precision per unit cost, relative to that corresponding to the optimal sampling fractions.

Table 6

Optimal Allocation of the Sample Across Strata for Overall Means, Given
Optimal Cluster Sizes in Each Stratum, for Various Relative Standard
Deviations Across Strata and Values of Intraclass Correlations

Ratios of Within Stratum Standard Deviations (High:Med:Low)	Ratios of Optimal Sampling Fractions (High:Med:Low)
$\rho = .005$	
3 : 2 : 1 1.7 : 1.5 : 1 1 : 1 : 1 .33 : .5 : 1	5.2 : 2.7 : 1 3 : 2 : 1 1.7 : 1.4 : 1 .6 : .9 : 1
$\rho = .01$	
3 : 2 : 1 1.7 : 1.5 : 1 1 : 1 : 1 .33 : .5 : 1	5.2 : 2.7 : 1 3 : 2 : 1 1.7 : 1.4 : 1 .6 : .9 : 1
$\rho = .02$	
3 : 2 : 1 1.8 : 1.5 : 1 1 : 1 : 1 .33 : .5 : 1	5.1 : 2.7 : 1 3 : 2 : 1 1.7 : 1.3 : 1 .6 : .9 : 1

6. SUMMARY

Rare population sampling forces the survey statistician to consider combinations of PSU and cluster definitions, stratification, and alterations of measures of size which are not typically found in cross-section samples. This research found that these traditional sample design techniques can be adapted to increase the efficiency of two-stage telephone samples for the Black household population with telephones.

First, this research found that even the rough correspondence between telephone exchanges and large cities and states permitted stratification that successfully discriminated exchange groups with vastly different eligibility rates. The high density stratum had over twice the proportion of Black households as did the low density stratum. This permits control over screening costs in sample implementation. With other rare populations which are residentially segregated, similar results are expected.

Second, the use of rejection rules based on subpopulation eligibility effectively reduced screening costs within PSUs. This increases the eligible proportion of secondary numbers from twofold to ninefold, depending on which density stratum was considered.

Third, use of a larger PSU (200- versus 100-series of consecutive numbers) produced no serious loss of eligibility. Hundred series densely filled with eligible numbers tend to be adjacent to others densely filled. This is a discovery concerning the practice of assigning numbers by telephone companies. This fact permits larger numbers of sample numbers per PSU, another key feature in reducing the costs of the Black population sample.

Despite great pressures for cost reduction in rare population samples, it is important to balance errors and costs explicitly in choosing the final design. In this research such cost

and sampling error modeling suggested that disproportionate allocation of the sample to Black-density strata is desirable. In addition, it is most efficient to select a relatively large set of secondary numbers per PSU. This set is sufficiently large that the 200- or 400-series PSU definition must be used.

Although we have applied this design only to the Black population, its performance should be similar for other residentially segregated populations. This includes income groups, certain occupational groups, and ethnic groups.

In addition, the discoveries of this research may also have implications for cross-section samples. Increasing the PSU size from 100 to 200 consecutive numbers may be advantageous in a two-stage RDD design for sampling the general telephone household population. The larger 200-series would provide twice as many numbers to select from and, as with the rare population, the proportion of eligible numbers would tend to be similar to that found in the 100-series. Therefore, given low intracluster correlation values, the cluster size of eligible numbers for a design could be set much closer to the optimal size. Because all PSUs selected would be able to support the chosen number of sample numbers, the achieved cluster size of eligible numbers should also be less variable over PSUs and therefore the impact of compensating weighting on the variance of estimates should not be great.

REFERENCES

- BLAIR, J., and R. CZAJA (1982). Locating a special population using random digit dialing. *Public Opinion Quarterly*, 46, 585-590.
- GROVES, R.M. (1978). An empirical comparison of two telephone sample designs. *Journal of Marketing Research*, 15, 622-631.
- GROVES, R.M., and KAHN, R.L. (1979). *Surveys By Telephone*. New York: Academic Press.
- KISH, L. (1976). "Optima and proxima in linear sample designs", *Journal of the Royal Statistical Society*, Ser. A, 139, 80-95.
- KISH, L. (1965). *Survey Sampling*. New York: John Wiley.
- THORNBERRY, O.T., and MASSEY, J.T. (1983). Coverage and response in random digit dialed national surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 654-659.
- WAKSBERG, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 40-46.

Comparing Telephone and Face-to-Face Interviewing in the United Kingdom

W.M. SYKES AND M. COLLINS¹

ABSTRACT

This paper presents results from methodological experiments comparing telephone and face-to-face interviewing in surveys of the general population. The relatively low level of telephone ownership in the United Kingdom, especially among the less privileged, argues the need for a dual-mode approach combining telephone interviews with face-to-face interviews for those without telephones. This approach depends on the absence of differential mode-effects on the answers obtained or on the ability to account for these effects when they occur.

KEY WORDS: Telephone interviewing; Dual-mode interviewing; Social surveys; Response rates; Data quality.

1. INTRODUCTION

The choice of a mode of data collection for a survey depends upon the availability of facts about the alternatives. In the U.K., such facts about telephone interviewing have just recently begun to emerge. The necessary comparisons between telephone interviewing and other data collection modes have been carried out only in the last two years. This delay is surprising given the lively debate about the merits and drawbacks of telephone interviewing and the attention which the issue has received in other countries.

Two studies conducted by the Survey Methods Centre at Social and Community Planning Research comparing telephone and face-to-face interviewing provide the focus for this paper. Carried out in 1983 and 1984, these studies examine some of the central issues: the public's willingness to take part in telephone surveys and the kind, quality and volume of data that can be collected. The studies are described in Section 2 and their results presented in Sections 3 and 4. Reference is also made to another British study – an experiment carried out in 1985 by the Market Research Development Fund – and to the larger volume of methodological research conducted in other countries, particularly the United States.

2. THE SCPR STUDIES

Our research program reflected telephone ownership which is low by North American standards: about 75% of households possessed telephones in 1983. Non-coverage is substantial and crucial, for social researchers, because of its bias towards less affluent sectors of British society. In this context, the main objective was to evaluate dual-mode interviewing, where telephone owners would be interviewed by telephone, and non-owners face-to-face.

The first study provided two comparisons towards this evaluation: between an experimental dual-mode sample and a larger national sample interviewed face-to-face; and between two samples of telephone owners, one sample interviewed by telephone, the other interviewed face-to-face. In this paper, we focus on the latter comparison, which addresses the question

¹ W.M. Sykes and M. Collins, Survey Methods Centre, Social and Community Planning Research, 35 Northampton Square, London EC1V 0AX, England

that lies at the heart of any evaluation of the dual-mode approach: are telephone and face-to-face data compatible or are there modal differences between them? If there are modal differences, the data cannot be “added” together and treated as a single data set without the kind of adjustments not usually possible in a one-time survey. The second study concentrated only on this direct comparison between the two interview methods among telephone owners.

2.1 Study 1

The first study was conducted alongside the 1983 British Social Attitudes Survey, which is here referred to as the “main” survey. This survey involved face-to-face interviews of about an hour, covering a wide range of political, economic, social and moral issues.

The sample for the main survey was about 1,750, and was representative of adults aged 18 or over living in private households. For practical reasons, the sample was confined to those at addresses in the Electoral Register. People living in institutions (though not private households at such institutions) were excluded, as were the 4% of adults known to live at addresses not on the Electoral Register (Todd and Butcher 1982).

A multi-stage design was used with four stages of selection: 103 constituencies in England and Wales and 11 local authority districts in Scotland were selected with probability proportional to electorate; within each a single polling district was selected, again with probability proportional to electorate; from each polling district, 23 addresses were selected with probability proportional to the number of electors registered at the address. At the final stage, one person at each address was selected by the interviewer, using an adaptation of the Marchant-Blyth procedure (Blyth and Marchant 1973).

For the experiment, a parallel sample of about 800 addresses (seven per area) was selected from the same 114 sampling points. These addresses, together with all the names in the Electoral Register, were submitted to British Telecom’s telephone number-retrieval facility. The facility yielded telephone numbers for 65% of the submitted addresses. Most of the difference between this retrieval rate and the level of telephone ownership – around 75% at the time – can be explained by ex-directory numbers: about 12% of telephone numbers in Great Britain are ex-directory, with regional and other variations as noted by Collins and Sykes (1987). Other problems in tracing telephone numbers seem to have had little effect.

The following procedure was used by British Telecom for retrieving telephone numbers: once the correct telephone exchange area had been identified by the address, the subscriber’s name was looked up in the directory. Specific address details (i.e., the street name) helped distinguish between subscribers with identical names. Since it is not clear from the Electoral Register which of the names at an address is that of the subscriber, British Telecom was asked to check every name before abandoning a search.

The telephone numbers obtained were systematically assigned to four sub-samples. Two of these were interviewed by telephone using a questionnaire expected to take about 20 minutes to complete. The questions were drawn from all sections of the main Social Attitudes questionnaire. The other two sub-samples were interviewed by telephone using a longer questionnaire – estimated at 40 minutes – that was also drawn from the main survey questionnaire. Sub-samples allocated to both the 20-minute and the 40-minute questionnaires were sent a letter before the telephone calls. The other sub-samples received no advance warning of the survey. In all cases the selection of a respondent for interview was on the same basis as for the main survey.

Experimental sample addresses for which no telephone numbers could be obtained from British Telecom were given face-to-face 20-minute interviews. Combined with those obtained by telephone, these interviews formed a dual-mode survey that was compared with the main face-to-face interview survey (Sykes and Hoinville 1985).

A more direct examination of interview mode effects was sought by submitting a systematic sub-sample of 600 of the main sample addresses (five in each area) to British Telecom's number-retrieval service. In this case, numbers were returned for 55% of the addresses (the variability of the success rate of the British Telecom number-retrieval service remains unexplained). Comparisons were then made between those who were interviewed by telephone and those who could have been interviewed by telephone but were interviewed face-to-face. By restricting comparisons to the telephone-accessible population, we controlled for effects attributable to differences between the compared populations rather than to differences in the mode of data collection.

2.2 Study 2

The second experiment concentrated on this direct comparison. About 2,300 addresses were selected from the Electoral Register, as in Study 1, and were sent to British Telecom for telephone numbers (with in this case, a 61% retrieval rate). Addresses for which telephone numbers were retrieved were split into three sub-samples. One group was interviewed by telephone using "pencil and paper" methods; another was interviewed using Computer Assisted Telephone Interviewing (CATI); the third was interviewed face-to-face. Our experiment with CATI was a practical failure (for a number of reasons), but the other two sub-samples again give us a direct comparison between people interviewed by telephone and people who could have been interviewed by telephone but were interviewed face-to-face. The questionnaire, designed to take 25 minutes, consisted of a sub-set of questions from the 1983 British Social Attitudes Survey.

2.3 Limitations on the Comparisons between Interviewing Modes

Three factors could limit comparisons between the answers obtained face-to-face and those obtained over the telephone. First, differential non-response (as discussed in Section 3) could have led to differences in the composition of the respondent sets. This possibility was tested using a number of demographic and socio-economic variables believed to be associated with certain attitude variables. Significant differences between the respondent sets suggest that, quite apart from any differences between the modes in overall response levels, certain kinds of people are more likely to participate in a telephone rather than a face-to-face survey, and vice versa. The variables examined were: age within sex, marital status, household composition, economic status, socio-economic group and geographical location. No statistically significant evidence of differential non-response was found in the first study. In the second study, two variables showed statistically significant differences between the telephone and face-to-face samples: household composition (the telephone respondents included a higher proportion of childless couples under 60, while the face-to-face sample had a higher percentage of couples with young children and teenagers); and socio-economic group (intermediate and junior non-manual workers and those in "other" occupations had greater representation in the telephone sample than face-to-face, and "homemakers" were a higher proportion of the face-to-face sample). These differences may well represent only sampling fluctuations, but they should lead to some caution in the interpretation of differences between the answers of the two samples.

The second possibility is of different levels of skill or supervision between the telephone and face-to-face interviewers. Six telephone interviewers were employed on the first experimental survey. Two were fully trained and experienced face-to-face interviewers, but the remainder had had no previous interviewing experience and so received basic interviewer training as well as the special telephone interviewing training that all six interviewers underwent. The second study involved 10 interviewers, three of whom had worked on the previous study. As in the previous study, a supervisor was present to listen in, advise on interviewing technique when necessary and check for obvious errors in completed questionnaires.

The face-to-face interviewers for both studies were drawn from Social and Community Planning Research's panel of about 300 regularly employed face-to-face interviewers. Their training in basic interviewing techniques was similar to that given to the telephone interviewers. However, for the most part, the face-to-face interviewers were more experienced than their telephone counterparts. Differences between the two groups of interviewers should, therefore, be kept in mind, especially differences suggesting lower quality in the telephone interviews.

The third factor is the questionnaires. The main Social Attitudes questionnaire, comprising about 100 questions, was divided into five broad topic areas: employment, education, health and housing, issues of social class, and racial and sexual equality. The experimental questionnaires were composed of those questions considered most important in the main survey. These questions were chosen to represent the full range of question types in the main questionnaire.

As a result, the experimental questionnaires covered a range of topics (including some "sensitive" issues) and included questions involving different kinds of response tasks and levels of complexity. The order of the questions on the Social Attitudes Survey was maintained for both the 20-minute and 40-minute experimental questionnaires used in the first study and for the 25-minute questionnaire used in the second study. Thus the 40-minute questionnaire was not made up of the short questionnaire followed by a further 20 minutes of questions: rather, questions from the 20-minute version were spread throughout. Alterations to question wording were made only when unavoidable; for example, re-wording to adjust for the necessary absence of showcards. The Social Attitudes Survey questionnaire consists largely of closed questions, so few of the results from our experiments relate to open questions.

All of these limitations should be kept in mind when examining our results, but they are largely inevitable in such comparative studies. As described above, we have tried to identify and minimize them. They are of great concern only when our results suggest mode effects that might confound the effects of other variables: most of our results do not point to this. Thus the limitations should be considered only as potential sources of effects counteracting mode effects we might otherwise have found – surely a less serious threat to the validity of our conclusions.

3. RESPONSE RATES

In the U.K., doubts about the feasibility of telephone interviewing, particularly for social surveys, stem from concerns not only with the level of communication possible, and its effect on both cognitive and affective dimensions of the interview, but also with the general social acceptability of this use of the telephone. In Britain, it is a common belief among researchers that "cold calls" from strangers are likely to be treated with circumspection: a call from a telephone interviewer may be regarded as inappropriate and intrusive.

A common counter argument points out the possible advantages telephone interviewing has over face-to-face interviewing, particularly in inner city areas. Escalating personal and property crime has led to increasing suspicion of strangers, which means falling response rates and the installation of devices such as entry-phones that make it harder for personal interviewers to contact respondents. By telephone, contact will also certainly be made at an address if someone is there, and, if not, subsequent attempts are not expensive.

Table 1 shows the response rates for both studies conducted by the Survey Methods Centre.

Table 1
SCPR Experiments: Response Rates

Bases	Study 1		Study 2	
	Telephone (429)	Face-to-Face (313)	Telephone (730)	Face-to-Face (631)
	%	%	%	%
Completed interviews	53	60	46	68
Partial interviews	1	-	-	-
Refusal (no selection)	5	2	21	6
Refusal (proxy)	9	5	7	4
Refusal (selected person)	11	18	10	11
No contact ^a	3	1	8	4
Selected person never in	3	3	3	2
Ill, away, language problems	2	5	2	4
Other ^b	13	6	4	2

Study 1: $\chi^2 = 3.72$ d.o.f. = 1 0.05 < p < 0.1

Study 2: $\chi^2 = 66.22$ d.o.f. = 1 p < 0.001

Studies 1 and 2 combined: $\chi^2 = 59.46$ d.o.f. = 1 p < 0.005

}

comparisons with only two categories:
completed interviews and
non-completed interviews.

^a Includes "Ring no answer" and "Permanently engaged".

^b Includes "Broken appointments", "Too old", "Incapacitated", "No connection", "Right number, wrong address".

Response to these studies, for both the telephone and face-to-face components, was relatively low. (We would normally expect personal interview response rates of over 70% before reissue of refusals.) This owes something to the nature of the surveys – general purpose surveys are notoriously difficult to “sell” to respondents. The same argument can also be applied to the only other major British methodological comparison survey, carried out by Marplan on behalf of the Market Research Development Fund. This study used the same sampling method as our own experiments and also included a wide range of general questions, under the title *Lifestyle in the 1980’s*. In this case, the response rates obtained were 45% by telephone with a sample base of 1697 and 67% face-to-face with a sample base of 1233 (Market Research Development Fund 1985). In both our studies, the response rate was lower for telephone interviews: barely half of the issued addresses yielded interviews. As Table 1 shows, the difference was on the borderline of non-significance for Study 1 but was statistically significant for Study 2 and for Studies 1 and 2 in combination.

The difference might be attributed to our relative lack of experience with telephone interviewing, but it is consistent with findings from other countries. For example, in the United States lower response rates – mostly arising from the higher incidence of refusals to cooperate – have been reported by a number of authors (e.g., Hochstim 1967; Henson, Roth and Cannell 1977). The position is summarized by Groves and Kahn, who write:

“The response rate of national surveys remains at least five percentage points lower than that expected in personal interview. This has been a rather stable comparison despite changes over time in training of interviewers, monitoring techniques, feedback procedures from monitors, and techniques of introducing the survey to the respondent.” (Groves and Kahn 1979; p. 219)

These findings suggest that sociological and psychological explanations of resistance to the telephone approach may be more appropriate than explanations of interviewer and general

methodological inexperience. However, the first SCPR study appears to have been rather more successful than either the second or the MRDF study. It has been suggested that this difference was due to the interest and excitement surrounding the first experiment. This may have communicated itself to the interviewers (for example, researchers were continually “dropping in” to observe the proceedings), thus affecting their success rates. Certainly, experience with face-to-face surveys suggests that interviewer morale and energy are important for good response rates.

In the SCPR studies two survey conditions were varied to assess their impact on telephone response rates. For the first survey, half the telephone respondents were asked to do 20-minute interviews and the other half did 40-minute interviews (respondents were told the length of the interview towards the end of the introduction), and in both surveys advance letters giving notice of the interview were sent to a random half of the telephone sample.

Table 2 shows that response for the 40-minute interview was lower than for the 20-minute interview, although the difference between the overall distributions was not significant. The main single reason for this lower response was the higher direct refusal rate, possibly indicating that respondents were less willing to undertake the longer interviews. However, very few respondents who had agreed to participate terminated an interview prematurely – even with the longer interview.

Different strategies may be needed for longer questionnaires. While it may be reasonable to request respondents to take part in a 20-minute interview at the time when first contact is made, a system of appointments may be more successful where more interviewing time is required. Wiseman and McDonald (1979) suggest that refusal rates are likely to be lower when interviewers are instructed to make call-back appointments should the respondents indicate that they are busy.

In other studies, sending advance letters to potential telephone respondents has been found to improve response rates. For example, Dillman, Gallegos and Frey (1976) obtained refusal rates which were, on average, 6% lower for respondents receiving advance letters (compared with 14%). As Table 3 shows, in the SCPR experiments response rates were slightly higher among respondents who had been sent an advance letter (no record was kept of whether letters had been received) although the differences were not statistically significant.

To explore why respondents refuse to be interviewed by telephone, 55 refusers to the first study were followed-up to see whether they would have co-operated at the first contact if they had been approached personally. Forty said that the method of interview would have made no difference to their decision, and only a very small number of these people subsequently agreed to be interviewed. Most of the rest said they would have taken part if they had been approached face-to-face and eventually completed a face-to-face interview (13 out of 15).

Because face-to-face refusers were not followed up, we do not know if a proportion of this group would have preferred to be approached by telephone.

3.1 Response Differences and Data Quality

The public's perception of the proper use of the household telephone may effect not only response rates, but also the kinds of questions respondents will be prepared to answer. Of even greater concern, however, is the type of communication possible between interviewer and respondent and its potential effect on the measurements made.

Face-to-face communication takes place both verbally and non-verbally, while the telephone has only limited channel capacity with exchanges between interviewer and respondent restricted to what is said and so-called paralinguistic cues: tone of voice, pauses and so on (Miller and Cannell 1982).

Table 2
SCPR Experiments: Effects of Interview Length (Study 1)

Bases	40-Minute (206)	20-Minute (223)
	%	%
Completed interviews	48	59
Refusal	27	23
Other	25	18

$\chi^2 = 4.7$ d.o.f. = 2 0.10 > p > 0.05

Table 3
SCPR Experiments: Effects of Advance Letters on Response Rates

Bases	Study 1		Study 2	
	Letter (215)	No Letter (214)	Letter (388)	No Letter (392)
	%	%	%	%
Completed interviews	55	51	48	43
Refusal	23	27	37	38
Other	22	21	15	19

Study 1: $\chi^2 = 1.09$ d.o.f. = 2 p > 0.5

Study 2: $\chi^2 = 2.8$ d.o.f. = 2 p > 0.2

Studies 1 and 2 combined: $\chi^2 = 3.49$ d.o.f. = 2 p > 0.1

The possible implications for survey measurements of the telephone’s limited channel capacity are numerous. For example, the absence of visual aids may increase the difficulty of some response tasks. “Voice only” communication may not convey the full meaning behind respondents’ words (making it difficult, for example, to probe open-ended questions) and may not reveal if they actually understand the questions. There may also be limitations on the interviewer’s ability to perform his or her role. Can verbal signals, for example, replace the non-verbal cues that convey interest and attention to the respondent, or those that help control the interview? Can the interviewer hold the concentration of the respondent, particularly in long interviews? Conversely, is the absence of visual stimuli a desirable reduction in the many sources of variability in survey data? Finally, does the greater social distance in the telephone interview make the respondent more or less comfortable in revealing sensitive information such as income, or information with a strong social desirability component?

SCPR’s experiments addressed some of these issues.

3.1.1 General Comparisons

Given the different refusal rates of the interviewing modes, it is surprising that there are few other general differences. This result has been replicated in many studies in the U.S. (Groves and Kahn 1979; Lucas and Adams 1977; Jordan *et al.* 1980; Colombotos 1969; Wiseman 1972), and in other countries such as Denmark (Kormendi *et al.* 1986). Simple straight-forward questions asked identically by telephone and face-to-face yield similar distributions of response.

In the SCPR studies the marginal distributions of response yielded by the different modes of interview were compared and differences were tested for statistical significance using chi-squared tests. These tests were performed on unweighted data. However, tables in the text, unless otherwise indicated, show distributions of data weighted to take account of any differences between the number of people listed on the Electoral Register and those found at an address. Such differences occurred in approximately 25% of cases, in each of which the data were weighted by the number of persons aged 18 or over living at the address divided by the number of electors listed on the Register for that address. Weighted tables are given to allow readers to decide if they might draw different conclusions from telephone survey data and face-to-face survey data when both sets have been prepared according to routine procedures.

Standard chi-squared tests were performed even though the data arose from a multi-stage sample. It has been shown (see, for example, Holt, Scott and Ewings 1980) that underestimating true variability by ignoring sample design will generally lead to test statistics which are too large, and hence to the false rejection of null hypotheses (i.e., to anti-conservative tests). For the Social Attitudes Survey, however, estimation of true standard errors for attitudinal variables yields Design Factors (the ratio of the complex standard error to the simple random sampling standard error) which are rarely above 1.2 (Jowell and Witherspoon 1985). Further, the literature argues that in 2-way tests of independence the consequences of clustering are likely to be less severe (Holt, Scott and Ewings 1980). As a result, we feel justified in using standard chi-squared tests to avoid the large amount of computation necessary for corrected statistics. If anything, this approach will overstate the significance of differences between interview modes.

In the first study we looked at 95 questions and parts of questions and in the second study 69. The results are shown in Table 4. It is clear that in both studies the results accorded with those of other researchers: the interviewing modes yield significantly different distributions of answers for only a very small percentage of questions. A similar finding emerged from the MRDF study.

3.1.2 Comparisons for Particular Question Forms

Despite the general result, research in the U.S. has shown that there are specific kinds of questions for which differences in response distributions do occur. For example, Groves and Kahn (1979) demonstrated a tendency for respondents to give truncated answers to open-ended items over the telephone. This might be due to the faster pace of telephone interviewing, as noted, for example, by Dillman (1970) and Williams (1977). Both interviewers and respondents tend to speak more quickly on the telephone and to avoid silent pauses. The swifter pace of telephone interviews was shown in our second experiment. As Table 5 shows, with an interview designed to take 25 minutes, 10% of the telephone interviews were conducted in under 20 minutes, compared with 5% of face-to-face interviews. At the other extreme, 41% of face-to-face interviews took more than half an hour compared with under a third of the telephone interviews.

Ball (1980) suggests that the greater speed may occur because the norms of telephone conversations require both the interviewer and respondent to work to maintain the conversational flow. This may leave respondents with less time to think about their answers. Certainly, silences seem to make people uncomfortable – in a study by Jordan (1980) routine pauses in the interview were described as interminable by interviewers. Undoubtedly there are many other contributing factors: even the absence of visual distractions may be important.

Although SCPR's experimental studies did not carry any open-ended items, the MRDF study included a number of spontaneous awareness measures. Comparisons of telephone and face-to-face results appear consistent with the findings discussed above. One example

Table 4
Differences in Marginal Distributions of Response:
Telephone vs. Face-to-Face

Bases	Study 1 (95)	Study 2 (69)
	%	%
No significant difference	91	87
Significant at 5%	7	9
Significant at 1%	2	4

Table 5
Interview Length by Mode of Interview (Study 2)

Unweighted Bases	Telephone (354)	Face-to-Face (360)
	%	%
Minutes		
Under 20	10	5
20-29	63	53
30-40	22	33
40 +	6	8

$\chi^2 = 17.6$ d.o.f. = 3 $p < 0.01$

Table 6
Comparisons of Responses on an Open Question (MRDF Survey)

What do you like about . . . soup?		
Bases	Telephone (700)	Face-to-Face (601)
	%	%
Number of answers		
None	33	22
One	58	61
Two	7	14
Three or more	1	2
Average	0.77	0.96

$\chi^2 = 32.2$ d.o.f. = 3 $p < 0.01$

is given in Table 6, which shows that a third of telephone respondents gave no answers, compared with under a quarter face-to-face. Also, the average number of responses given over the telephone was significantly lower.

We might assume that more or longer answers mean more valid reporting, and this would imply a need for techniques to improve open questions on telephone surveys. At the extreme, it might be concluded that open questions have only limited use on telephone surveys, for example when only the first information spontaneously offered by respondents is wanted. This assumption needs, however, to be tested: here we can only report the effect.

Differences between response distributions have also been reported for attitude scale questions asked identically face-to-face and over the telephone. Telephone respondents tend towards "acquiescence" and "extremeness" response bias (Jordan, Marcus and Reeder 1980; Groves and Kahn 1979). With the agree/disagree scales used by MRDF, the telephone sample showed a slight tendency to agree more. However, no difference in the spread of responses was found – there was no evidence of a greater tendency towards extremeness.

3.1.3 Sensitive Questions

Concerning the types of question that can be used in telephone surveys, researchers have paid much attention to sensitive questions – those that deal with private or personal information and those for which certain responses are more clearly socially acceptable. Initial views about the likely effects of asking sensitive questions over the telephone were divided. Those who felt that respondents would be less willing to answer truthfully said that the lack of the interviewers' reassuring presence would make respondents less likely to be frank and open. The opposite view – that respondents would give more valid answers – maintained that greater social distance, by preserving anonymity, would encourage truthful responses.

Most evidence supports the latter view (Colombotos 1965; Wiseman 1972; Henson, Roth and Cannell 1974; Locander 1974; Rogers 1976). The major exception is reported by Groves and Kahn (1979), who found telephone respondents to be reticent about their financial status and other sensitive issues.

Our studies support the hypothesis that telephone surveys work well for sensitive questions. For instance, in our first study 14 questions were isolated as potentially sensitive and tested for mode-effects. Three illustrative examples of such questions are given below:

i) How would you describe yourself?:

(Read out) ...

... as very prejudiced against people of other races

... a little prejudiced

... or, not at all prejudiced?

ii) Do you think, on the whole, that Britain gives too little or too much help to Asians and West Indians who have settled in this country, or are present arrangements about right?

iii) Finally in this section, I would like you to tell me whether, in your opinion, it is acceptable for a homosexual person to be a teacher in a school?

No significant differences in the marginal distributions of response were found. For several questions, however, there was a somewhat greater tendency to give socially desirable answers in face-to-face contact. In other words, the questions seemed to be less sensitive over the telephone. For example, 28% of respondents interviewed by telephone admitted to having been questioned by police over the past two years in connection with a crime, compared with 20% of face-to-face respondents.

Sensitive questions in the MRDF study also showed a slight tendency for telephone respondents to give more “honest” answers, although on individual questions differences in the distributions were generally not significant. For example, when asked to describe themselves on a number of dimensions, telephone respondents were more likely to say they were “attractive” (mean score of 2.81 out of 4 compared with 2.72 face-to-face) and were more ready to give an answer at all (88% gave an answer compared with 75% face-to-face).

Questions about income have generally been regarded as potentially problematic in telephone surveys, both in respondents’ willingness to answer and in the answers given. Under-reporting of income levels is the main expectation, although in practice this may be hard to distinguish from under-estimation resulting from higher non-response in the upper income brackets. A study by Locander and Burton (1976) suggests that the validity of income data may depend on the question format. In a comparison of four question formats, under-reporting of income resulted from a method that first asked “Is your income more than \$2,000?” gradually increasing the figure until the first “no” response. However, over-reporting of income was encouraged by a similar method that began with the highest income category. The method used for the telephone surveys in the SCPR experiments was similar to the first type described above. It most closely approximates the response task set by the face-to-face income question in which a card indicating broad income bands, starting with the lowest, was used to guide the respondents’ choice. Over the telephone, the ranges were read to respondents starting at the lowest levels. The results are shown in Table 7.

In neither study was there any mode difference in respondents’ willingness to answer the income question. Differences in the distribution of answers, in this case a possible under-reporting of income, were only apparent in the first study.

3.1.4 Complex Questions

In both SCPR studies a number of questions were identified in advance as likely to pose particular response problems for telephone respondents. These included questions with one or more potentially difficult concepts, long questions and questions with large numbers of response options. Such “complex” questions appear to be no more problematic for telephone respondents than for those interviewed in person. For example, of 19 “complex” questions

Table 7
Gross Household Income: SCPR Studies

Bases ^a Income	Study 1		Study 2	
	Telephone (183)	Face-to-Face (170)	Telephone (297)	Face-to-Face (352)
less than £5,000	38	27	28	28
£5,000-£9,999	42	37	37	38
£10,000 or over	21	35	35	35

Study 1: $\chi^2 = 10.08$ d.o.f. = 2 $p < 0.01$

Study 2: $\chi^2 = 0.11$ d.o.f. = 2 $p > 0.9$

Bases	(217)	(199)	(344)	(405)
Don't know/ Not answered	16%	15%	14%	13%

^a “Don’t know” and “Not answered” excluded.

identified on the first study (12 of which had been asked with the aid of show-cards face-to-face), only one showed any evidence of mode-effects.

4. SUMMARY AND CONCLUSIONS

Since telephone ownership in the United Kingdom remains relatively low, particularly for certain sectors of the population, telephone interviewing is unlikely to replace face-to-face interviewing for surveys that must include the less advantaged. But its potential in combination with traditional face-to-face procedures has gained recognition. For example, the U.K. Labour Force Survey uses telephone interviewing for second and subsequent interviews with eligible respondents who have indicated a willingness to be contacted by telephone.

Crucial to the success of dual-mode surveys is the absence of differential mode effects. The results reported here provide a largely optimistic outlook. With a few exceptions there were no statistically significant differences between the distributions of answers obtained face-to-face and those given over the telephone.

However, the relatively low response rates to telephone surveys poses problems that need to be overcome. High refusal rates can reduce the cost-effectiveness of using the telephone. More importantly, they increase the chances of introducing bias into the sample. Further research to explore ways of improving telephone response rates is necessary to realize the potential of the method in the United Kingdom.

ACKNOWLEDGEMENTS

This research was carried out within the SCPR Survey Methods Centre, funded as a Designated Research Center by the Economic and Social Research Council (grant HR 3333). The authors are grateful to the editors and referees for their extensive and useful comments on earlier drafts of this paper.

REFERENCES

- ARONSON, S. (1971). The Sociology of the Telephone. *International Journal of Comparative Sociology*, 12, 153-167.
- BALL, D.W. (1968). Towards a Sociology of Telephones and Telephoners. In *Sociology and Everyday Life* (ed. Marcello Truzzi), Englewood Cliffs, New Jersey: Prentice Hall.
- BERGSTEN, J.W. (1979). Some Methodological Results from Four Statewide Telephone Surveys Using Random Digit Dialing. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 239-243.
- BISHOP, Y.M.N., FIENBERG, S.E., and HOLLAND, P.W. (1975). *Discrete Multi-variate Analysis: Theory and Practice*. Cambridge, Mass: MIT Press.
- BLANKENSHIP, A.B. (1977). *Professional Telephone Surveys*. London: McGraw-Hill.
- BLYTH, W.G., and MARCHANT, L.J. (1973). A self-weighting random sampling technique. *Journal of the Market Research Society*, 15, 157-162.
- CANNELL, C.F., OKSENBERG, L., and CONVERSE, J.M. (1979). Experiments in interviewing techniques. Research Report, Institute for Social Research, University of Michigan.
- CHRISTOFFERSEN, M.N. (1984). The quality of data collected at telephone interviews. Danish National Institute of Social Research, Copenhagen.

- COLLINS, M. (1983). Telephone interviewing in consumer surveys. *Market Research Society Newsletter*, October.
- COLLINS, M., and SYKES, W. (1987). The Problems of Non-Coverage and Unlisted Numbers in Telephone Surveys in Britain. *Journal of the Royal Statistical Society*. Ser. A, 150, (forthcoming).
- COLOMBOTOS, J. (1965). The effects of personal vs. telephone interviews on socially acceptable responses. *Public Opinion Quarterly*, 29, 457-458.
- COLOMBOTOS, J. (1969). Personal versus telephone interviews: effect on responses. *Public Health Reports*, 84, 773-782.
- COOMBS, L., and FREEMAN, R. (1986). Use of telephone interviews in a longitudinal fertility study. *Public Opinion Quarterly*, 28, 112-117.
- CZAJA, R., BLAIR, J., and SEBESTIK, J.P. (1982). Respondent selection in a telephone survey: a comparison of three techniques. *Journal of Marketing Research*, 19, 381-385.
- DE MAIO, T.J. (1984). Refusals in telephone surveys: when do they occur? Paper presented at the 39th Annual Conference of the American Association for Public Opinion Research.
- DILLMAN, D. (1970). *Mail and Telephone Surveys: The Total Design Method*. New York: John Wiley.
- DILLMAN, D., GALLEGOS, J., and FREY, J. (1976). Reducing refusal rates for telephone interviews. *Public Opinion Quarterly*, 40, 66-78.
- FALTHZIK, A. (1972). When to make telephone interviews. *Journal of Marketing Research*, 9, 451-452.
- FITTI, J.E. (1979). Some results from the telephone health interview survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 244-249.
- FLEISHMAN, E., and BERK, M. (1979). Survey of interviewer attitudes towards methodological issues in the national medical care expenditure survey. Paper presented at the Third Biennial Conference on Health Survey Research and Methods, Reston, Virginia.
- GROVES, R., and KAHN, R. (1979). *Surveys by Telephone: A National Comparison with Personal Interviews*. New York: Academic Press.
- GROVES, R.M., MAGILAVY, L.J., and MATHIOWETZ, N.A. (1981). The process of interviewer variability. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 438-443.
- GROVES, R.M., and MATHIOWETZ, N.A. (1984). Computer assisted telephone interviewing: effects on interviewers and respondents. Survey Research Center, University of Michigan.
- HENSON, R., ROTH, A., and CANNELL, C.F. (1974). Personal vs. telephone interviews and the effects of telephone re-interviews on reporting of psychiatric symptomatology. Research Report, Survey Research Center, University of Michigan.
- HOCHSTIM, J.R. (1967). A critical comparison of three strategies of collecting data from households. *Journal of the American Statistical Association*, 62, 976-986.
- HOLT, D., SCOTT, A.J., and EWINGS, P.D. (1980). Chi-squared tests with survey data. *Journal of the Royal Statistical Society*, Ser. A, 143, 303-320.
- IBSEN, C.A., and BALLWEG, J. (1974). Telephone interviews in social research: some methodological considerations. *Quality and Quantity*, 7, 181-192.
- JOWELL, R., and WITHERSPOON, S. (1985). *British Social Attitudes: The 1985 Report*. Aldershot: Gower.
- JORDAN, L.A., MARCUS, A.C., and REEDER, L.G. (1980). Response styles in telephone and household interviewing: a field experiment. *Public Opinion Quarterly*, 44, 210-222.
- KAHN, R.L., and GROVES, R.M. (1977). *Comparing telephone and personal interview systems*. Survey Research Center, University of Michigan.
- KORMENDI, E., EGSMOSE, L., and NOORDHOEK, J. (1986). Datakvalitet ved Telefon-interview. Socialforskningsinstituttet, Studie 52, Copenhagen.

- LOCANDER, W.B., and BURTON, J.P. (1976). The effect of question form on gathering income data by telephone. *Journal of Marketing Research*, 13, 189-192.
- LOCANDER, W.B., SUDMAN, S., and BRADBURN, N. (1974). An investigation of interview method, threat and response distortion. *Proceedings of the Social Statistics Section, American Statistical Association*, 21-27.
- LUCAS, W.A., and ADAMS, W.C. (1977). *An Assessment of Telephone Survey Methods*. Santa Monica, California: Rand Corporation.
- McCULLAGH, P., and NELDER, J.A. (1983). *Generalised Linear Models*. London: Chapman and Hall.
- MILLER, P.V., and CANNELL, C.F. (1982). A study of experimental techniques for telephone interviewing. *Public Opinion Quarterly*, 46, 250-269.
- Market Research Development Fund, (1985). *Comparing telephone and face-to-face surveys*. Marplan Ltd.
- OKSENBERG, L., COLEMAN, L., and CANNELL, C. (1984). Voices and refusal rates in telephone surveys. Unpublished manuscript.
- O'NEIL, M., GROVES, R., and CANNELL, C. (1979). Telephone interview introductions and refusal rates: experiments in increasing respondent cooperation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 252-255.
- ROGERS, T.F. (1976). Interviews by telephone and in person. *Public Opinion Quarterly*, 40, 51-65.
- SCHMIEDESKAMP, J.W. (1962). Reinterviews by telephone. *Journal of Marketing*, 26, 28-34.
- SYKES, W., and HOINVILLE, G. (1985). Telephone interviewing on a survey of social attitudes: a comparison with face-to-face procedures. Social and Community Planning Research Center.
- TODD, J., and BUTCHER, R. (1982). *Electoral Registration in 1981*. London: OPCS.
- WILLIAMS, E. (1977). Experimental comparisons of face-to-face and mediated communication. *Psychological Bulletin*, 84, 963-976.
- WISEMAN, F. (1972). Methodological bias in public opinion surveys. *Public Opinion Quarterly*, 34, 105-108.
- WISEMAN, F., and McDONALD, P. (1979). Noncontact and refusal rates in consumer telephone surveys. *Journal of Marketing Research*, 16, 478-484.

Issues in the Use of Administrative Records for Statistical Purposes

G.J. BRACKSTONE¹

ABSTRACT

Demands for statistics on all aspects of our lives, our society and our economy continue to grow. At the same time statistical agencies share with many respondents a growing concern over the mounting burden of response to surveys. One result of the search for alternative methods of satisfying statistical demands has been an increased emphasis on the use of administrative records for statistical purposes. This paper reviews recent experience at Statistics Canada in this area and discusses obstacles to the greater use of administrative records. Approaches to rendering administrative systems more useful for statistical purposes are reviewed, together with some important concerns related to information protection and record linkage.

KEY WORDS: Indirect estimation; Survey frames; Survey evaluation; Access; Confidentiality.

1. INTRODUCTION

Demands for statistics on many aspects of our lives, our society, our economy and our environment continue to grow. This may be due in part to our increased ability to handle and manipulate large sets of data as we move into the so-called information age, and it may also be a reflection of the increasing complexity of our social and economic systems and our desire to understand them better. Whatever their cause we face these demands in a climate of tight budgetary constraint for government statistical agencies. At the same time, statistical agencies are sensitive to the increased burden that would be imposed on respondents by an increase in survey-taking activity to meet these demands.

These factors have led to the exploration of other means of satisfying these statistical demands. Prominent among these alternative means is the increased use of existing administrative systems as sources of statistical data. This is not a new idea. For many years, statistical data have been a by-product of administrative processes in domains such as vital statistics, imports and exports, health care, and education. We will describe later how this usage of administrative data has spread more recently to statistics on businesses and on families and individuals.

The first sections of the paper describe the variety of types and uses of administrative records, illustrating some of their uses in Statistics Canada's program. The heavy dependency of Canada's statistical system on administrative records will be apparent. Section 6 discusses issues of accessing administrative sources and making them more appropriate for statistical use. Finally, a brief review of privacy concerns related to administrative record use is provided.

2. TYPES OF ADMINISTRATIVE RECORD

Administrative records come in many shapes and sizes. An important distinction is between those administered nationally (usually by the Federal Government) and those

¹ G.J. Brackstone, Assistant Chief Statistician, Statistics Canada, 26-J R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario K1A 0T6.

administered sub-nationally (e.g., by provinces or municipalities). For the latter to be useful nationally, agreement between jurisdictions is required on items such as definitions, standards, record formats, and procedures. Such agreement is not always easy to achieve, particularly in domains that are constitutionally within provincial jurisdictions.

Administrative records vary in terms of their purpose, and their purpose is a prime determinant of their coverage and quality, and therefore of their statistical usefulness. Six broad categories of purpose can be distinguished.

(1) *Records maintained to regulate the flow of goods and people across borders.*

These include records of imports, exports, immigration and emigration. The coverage and content of the resulting administrative records depend on the particular laws and regulations to be enforced, and on the success of their enforcement. Typically such laws are well enforced. Immigration records, by definition, exclude illegal immigrants but otherwise are complete. However, since emigration from Canada is not controlled, no direct administrative emigration records exist. Administrative records on Canadian imports tend to be more accurate than those on exports since the former require more detailed documentation in order to assess their liability for duty.

(2) *Records resulting from legal requirements to register particular events.*

Examples include births, deaths, marriages, divorces, business incorporations or amalgamations, licensing, etc. Typically coverage and quality of records collected for this purpose are very high in Canada, since evidence of this type of registration is necessary to obtain rights or benefits.

(3) *Records needed to administer benefits or obligations.*

Examples include taxation, unemployment insurance, pensions, health insurance, and family allowances. The coverage and content of these records are highly program dependent. The population to which they apply may be very well covered, but for political or administrative reasons the definition of this population may not be the most useful definition analytically.

(4) *Records needed to administer public institutions.*

These include, for example, records related to schools, universities, health institutions, courts and prisons. Such records tend to focus on the institutional caseload rather than on the individuals passing through the institution. On the other hand, they usually provide very complete aggregate statistics on the population using these institutions. In Canada, many administrative records in this category fall within provincial jurisdiction.

(5) *Records arising from the government regulation of industry.*

Examples include records in the areas of transportation, banking, broadcasting and telecommunications. They also include records arising from the management of the supply or the price of some commodities, especially in the agriculture area.

(6) *Records arising from the provision of utilities.*

These include electricity, phone and water services. Their coverage of subscribers and the quality of information associated with services and billing are normally good. Many of these services are administered at the provincial or municipal levels.

Administrative records also vary in terms of the processes by which they are assembled. Most administrative processes with wide coverage are now automated, but differences in hardware and data formats (both between jurisdictions, and between the administrative agency and the statistical agency) have to be faced. Increased automation also leads to an increasing amount of modification to the originally reported records by the administrative agency before they are received by the statistical agency. While enhanced control of the quality of incoming forms may be beneficial to the final quality of the administrative file, additional work is required by the statistical agency to understand and evaluate the effects of any preliminary processing carried out by the administrative agency. In some administrative systems, the individual records remain at their local source and only aggregates are assembled centrally. This practice restricts the statistical agency's ability to evaluate the quality of the data and limits flexibility in statistical analysis of the data.

Finally, records differ in terms of their accessibility. Legal and regulatory provisions often govern access to, and use of, administrative records for secondary, including statistical, purposes. This topic is addressed further in Section 6.

3. USES OF ADMINISTRATIVE RECORDS

The statistical uses of administrative records may be categorized into four main areas. Most statistical applications of administrative records fall into one of these four categories or represent combinations or variations of these uses.

(1) *Direct Tabulation*

This includes the counting of units in files, cross-classification by attribute, and the aggregation of quantitative variables associated with each unit. Statistics on vital events and on external trade are important examples. Other examples include the publication of monthly counts of unemployment insurance claimants, and of beneficiaries by province, age, sex and length and type of benefit, and annual summaries of income distributions for each county based on the personal income tax file.

(2) *Indirect Estimation*

This category includes cases where data from administrative records comprise one of the inputs into an estimation process. For example, individual tax returns for the same taxfiler are linked from one year to the next in order to produce partial estimates of migration which can be weighted up with reference to census-based benchmarks. These estimates of migration then feed into Statistics Canada's population estimation program (which also makes use of administrative data on births, deaths and immigration). A second example is the use of taxation data for small businesses in lieu of seeking survey data from them. These tax-based data, adjusted if necessary, are combined with survey-based data for large businesses to provide industry aggregates.

Also within this category are uses that involve the linkage of different administrative or statistical files to produce estimates. For example, the linkage of the death register with files of individuals exposed to particular hazards in order to estimate differential mortality rates, or the linkage of records from tax files, unemployment insurance files, and manpower training files in order to analyse labour market attachment and adjustment.

(3) *Survey Frames*

In this category we include the use of administrative records to create, supplement or update frames to be used for censuses or surveys. A primary example is the use of payroll deduction information submitted by employers to Revenue Canada. The questionnaire which has to be completed by new payroll deduction account holders is a valuable means of identifying new businesses or changes in the structure of existing ones. Although in Canada we do not have a register of housing units, a second example would be the use of building permits or new telephone or electricity connections as signals of possible new housing units.

(4) *Survey Evaluation*

This category covers the use of administrative records for checking, validating or evaluating survey-derived data. This may be done either at the individual unit level, or at an aggregate level. Several census evaluation studies in the past have used immigration and taxation records to evaluate census questions on immigration and income, respectively, while family allowance records have been used in checking the census coverage of children.

An important determinant of how a particular administrative source will be used is the perceived quality of the administrative records compared to corresponding survey information. In some instances administrative records are used to evaluate survey responses, while in others survey-based data provide a means of benchmarking administrative-based estimates. The quality of administrative records has to be assessed in each individual case. In general, their quality for statistical purposes depends upon at least three factors:

- (i) the definitions used within the administrative system;
- (ii) the intended coverage of the administrative system;
- (iii) the quality with which data are reported and processed in the administrative system.

Weaknesses in any of these three factors can affect the statistical usefulness of the administrative records. The timeliness with which they are available is also an important consideration. Some of the potential limitations that need to be considered when deciding on the statistical use of administrative records have been described elsewhere (e.g., see Brackstone 1984). The strengths and weaknesses of administrative records compared to those of censuses and surveys are summarized in Table 1.

To illustrate the utilization of administrative records in Canada we will describe two areas of application within Statistics Canada. The first deals with the production of business statistics; the second addresses the production of statistics on individuals and families.

4. ADMINISTRATIVE DATA AND BUSINESS SURVEYS

Statistics Canada is currently in the throes of a complete redesign of the infrastructure and strategy on which its business surveys program is based. In particular this involves the redesign of the business register (the frame for business surveys), the re-thinking of the role

Table 1

Comparison of Censuses, Surveys and Administrative Records as Sources of Statistical Data

Factors	Censuses	Surveys	Administrative records
1. Coverage	Aim at complete coverage of the population	Some surveys exclude certain sectors of the population (e.g., Indian reserves, remote areas)	Target populations are defined by administrative requirements
2. Content	Wide range of data items allows extensive cross-classification	Usually covers a narrow range of topics but in more depth than a census	Restricted to variables required for administrative purposes
3. Concepts/definitions	Can be based on the requirements of social and economic analysis	Can be based on the requirements of social and economic analysis	Defined by administrative requirements
4. Small area estimates	Available as a result of aim at complete coverage	Unavailable in most cases	Available, provided individual records are geographically coded to small areas
5. Quality control	Can be designed to minimize errors	Smaller size allows for even tighter control than in censuses	Under the control of the administrative agency and may not receive attention except for key variables
6. Cost	Expensive	Relatively low cost per survey, although the cumulative cost of a regular survey over a 5-year inter-censal period may be large	Relatively inexpensive if initial collection costs attributed to the administrative program
7. Frequency	Every 5 or 10 years (depending on topic)	May be annual, quarterly or monthly depending on topic	May be annual or monthly depending on administrative program
8. Timeliness	Data available six months to 2½ years after Census Day	Repeated regular surveys produce results in a few weeks. Ad hoc surveys may require several months	Dependent upon the administrative process. An annual file may not be available in a clean form until well into the following year
9. Stability	Changes are under the control of statisticians who respond to user needs	In repeated surveys, changes are infrequent to allow comparisons over time	Changes may occur due to legislative or regulatory change, or due to changes in administrative practice
10. Respondent burden	Heavy but infrequent. Reduced through the use of sampling	Light on average, though heavy for those selected	No additional burden

and use of tax data within the program, and the development of a consistent strategy for the design of both annual and sub-annual business surveys. This redesign was motivated by needs to:

- (a) overcome some noticeable data quality weaknesses in the current program;
- (b) better integrate data from different surveys;
- (c) minimize respondent burden by making maximum use of tax data;
- (d) reduce resources required for maintaining survey frames.

A more detailed description of this project can be found in Colledge (1987).

Income tax and payroll deduction data play a prominent role in the conduct of business surveys. Annual tax returns submitted by corporations (T2) and by individuals (T1) are available to Statistics Canada under the Statistics Act. The payroll deductions of income taxes by employers are also available. Statistics Canada makes use of these data from business for two distinct purposes:

- (i) maintenance of its frame of businesses;
- (ii) substituting income tax data for survey data.

4.1 Frame Maintenance

The maintenance of a frame of businesses is a complex task. This complexity stems primarily from the complex structure and inter-relationships of many businesses, particularly large ones, and from the difficulty of keeping track of the very large number of births and deaths occurring among small business. The term “business” itself needs careful definition. In fact a distinction must be made between legal structures (incorporated companies, etc.), operating structures (the way companies organize and operate themselves), and statistical structures (the units for which data are required for analytic purposes). A hierarchy of units can be defined within each of these structures. In the case of the statistical structure, Statistics Canada has defined a hierarchy comprising, from top down, enterprises, statistical companies, establishments and locations. The task of frame maintenance thus involves not only updating for births and deaths but also keeping track of changes in the relationships between the various units within complex businesses, including the relationships between the statistical and operating hierarchies.

The proposed frame strategy calls for the continuous maintenance of the current corporate structure of all companies above a certain threshold size (which varies with industry), including the relationship of this structure to tax reporting units. Companies updated in this way will account for at least 70% of economic activity in each industry.

An activity known as “profiling” is used to determine the internal structure of complex businesses. This involves interviewing officers of the business to understand their operating structure and identify the appropriate statistical units. An important source of information on changes to business (births and restructuring) is Revenue Canada’s payroll deduction (PD) system. The activation of a new PD account by an employer is treated as a signal that something has happened. Such signals are followed up with the business to identify whether a frame update is required. Other signals will be obtained from annual tax returns, from responses to regular surveys, and from routine profiling.

In the case of smaller companies, where the structure is usually simpler but the turnover is faster, no attempt is made to define the various types of unit and their inter-relationships.

Instead, administrative data are used directly. Two alternative lists of businesses are made available as a basis for surveying – one is the most recent set of annual tax returns; the second is the current set of PD accounts. In both cases, all units above the threshold are removed. These two lists overlap and the most appropriate one is used in each particular survey. The PD-based list, which is more current since PD accounts may be opened or closed at any time during the year, is preferred for sub-annual surveys. It has the disadvantage of excluding non-employers.

4.2 Substituting for Survey Data

In the interests of minimizing both response burden and costs, tax data are used to replace survey data where feasible. The concepts and definitions underlying tax data do not uniformly coincide with the survey definitions required to assure consistency in the System of National Accounts or for other analytic purposes. Therefore care has to be taken in selecting from tax returns the data items that come closest to the required survey definitions. Furthermore, tax data do not contain the full range of variables required by many annual business surveys. In particular, they lack production statistics.

A further problem in utilizing tax data lies in establishing the relationship between the unit for which a tax return is submitted and the unit(s) to be surveyed. This is a problem particularly for the large complex businesses referred to earlier.

The strategy that has been developed for annual surveys is to make use of tax data primarily for small businesses where there is usually a one-to-one relationship between the taxfiler and the business. This approach significantly reduces the response burden on small businesses, without unduly affecting the quality of final data, since the bulk of economic activity is reported through the survey returns of larger companies.

It is clear from this brief overview of the new business survey strategy and infrastructure that there is a fundamental dependence on tax data for the continuing functioning of the program. This requires a very close working relationship between Statistics Canada and Revenue Canada so that the impact of administrative and procedural changes in the tax system can be assessed and prepared for in advance.

5. SOCIO-ECONOMIC DATA FROM ADMINISTRATIVE SOURCES

A systematic effort to develop data on individuals, families and households from administrative records was initiated in the late 1970s. The original motivation for this work was the rising costs of census-taking and the search for cheaper alternatives. It quickly became apparent that the statistical potential of administrative records on individuals in Canada lay in supplementing the quinquennial census through the provision of data for small areas inter-censally, rather than in replacing the census. It is not possible to achieve the coverage, geographic specificity, and range of individual, family and household characteristics required from a census with the existing administrative record systems. Nevertheless, the emulation of census coverage using a combination of administrative record systems is being pursued, together with the study of the possibility of replacing some census questions with data derived from administrative sources.

This section will concentrate on the use of administrative records to supplement census data inter-censally. The focus of the developmental work has been on administrative record systems that are national in scope (e.g., income tax, unemployment insurance (UI), family allowance, old age security) rather than systems that are administered at provincial or lower

levels (e.g., health insurance, driver's licences, municipal assessments). In the latter case the problem of standardization across jurisdictions is added to the other problems inherent in the statistical use of administrative records.

The annual individual tax file (T1) has proven to be the principal source of statistical data on individuals. The first use of this file was its direct tabulation to produce statistics on income and labour force participants by age and sex for provincial and sub-provincial areas. Identification of geographic location of taxfilers is based on the postal code indicated on the record. A file that provides a conversion from postal code to the various levels of census geography (province, county, municipality, electoral district, etc.) has been developed. Special tabulations can also be produced for user-defined areas described in terms of postal codes.

Data derived in this way are, of course, based on the concepts, definitions and regulations implicit in the Income Tax Act. These may not conform to definitions desired for analytic purposes (e.g., some forms of social assistance which are not taxable may be excluded). Income can be broken down by source – in particular, employment income can be separated. Variables available for cross-classification are limited (e.g., age, sex and marital status). Occupation, though asked on the tax form, is not reported nor coded with sufficient quality to be statistically useful. The coverage of these data is limited by the need to file a tax return. Low income individuals and dependents are therefore under-represented. Over time, changes to tax law can have a significant impact on coverage; e.g., the introduction of the Child Tax Credit, that required low income earners to file a tax return in order to claim the credit, led to a marked increase in coverage in 1978 compared to the previous year.

Despite these reservations, data produced by direct tabulation from income tax files provide a useful inter-censal source of small area income data. A recent publication from Statistics Canada made use of this source to produce data for Forward Sortation Areas, i.e., the first three characters of the postal code (Statistics Canada 1987). Since a prime concern in the publication of data for small areas is to ensure that no individual data can be deduced from aggregate totals for small areas, data are not provided for areas with less than 100 taxfilers.

A second use of the individual tax file is for estimating annual migration. This is achieved by matching individuals on tax files for two successive years and comparing the Census Division (or county) code assignment for each year. If there has been a change in code, it is assumed that the taxfiler has migrated. Demographic and tax exemption information are used to estimate the total number of persons who have migrated with the taxfiler. In a final stage, since the tax file does not cover the whole population, an adjustment is made to estimate the total number of migrants from year to year. Since 1981, tax-based migration estimates have been used in Statistics Canada's population estimates program. A full description of the methodology for estimating migration from tax records can be found in Norris and Standish (1983).

While data on individual incomes can be derived from tax data as described earlier, more analytic and policy interest focuses on family income. To derive family income from the individual tax file requires the capacity to identify and match records of individuals belonging to the same family. Development of family income data in this way has been proceeding with encouraging results. A description of methodology and results can be found in Auger (1987).

A second important administrative source of data on individuals is the unemployment insurance (UI) system. Files of both claimants and beneficiaries are available to Statistics Canada. The UI claimant and beneficiary files contain individuals who, for a variety of reasons, may be entitled to UI benefits. Not all of these individuals are considered to be unemployed according to the standard international definition of unemployment as incorporated in the Labour Force Survey (LFS), the source of published unemployment rates.

If a closely corresponding category in the UI system can be found, these files can be used to tabulate counts of "unemployed" for small areas. However, since even the best choice of category in the UI system does not correspond exactly with the definition of "unemployed" used in the LFS, attention has to be focused on how to integrate or reconcile these two sources of data. For example, monthly counts for small areas from the UI system might be used as indicators of changes in unemployment at the local level which could be calibrated to reliable LFS estimates at a higher geographic level (e.g., the province). Various methods of estimation along these lines have been investigated (e.g., regression estimation, SPREE - structure preserving ratio estimation), though without as yet any final conclusion as to the most appropriate method. A description of this work can be found in Trotter and Choudhry (1985) while Feeney (1987) describes a similar approach in the Australian context. A time series modelling approach which exploits the correlated structure of the error over time appears very promising (Choudhry and Hidirolou 1987).

These examples have illustrated that, in the case of statistics on individuals, the primary uses of administrative records are for direct tabulation and as input into estimation processes. This contrasts with the examples from the business side where frame maintenance and substitution for survey responses were the main uses.

While these two examples represent two important developing areas of administrative record use in Statistics Canada, they cover only a small fraction of the administrative files used by the Agency. There is, for example, a widespread and long-standing use of administrative records in the social institutions area (education, health, justice) both for creating survey frames and for obtaining statistical data. Current developmental work on telephone surveying and on address registers is using administrative records to develop frames of dwellings or households. A recent internal survey identified more than 50 administrative systems being used for statistical purposes. These covered the full range of types and uses described in Sections 2 and 3, and included examples from areas as varied as disease registries, motor vehicle licences, aircraft landings, milk marketing boards, fuel sales tax, municipal construction records, and customs and excise.

6. ACCESSING AND INFLUENCING ADMINISTRATIVE SYSTEMS

It is clear from this review of the use of administrative records for statistical purposes, that administrative records are a vital input to many of Statistics Canada's programs. This leads to a consideration of measures the Agency can take to protect the supply of data from administrative sources, and perhaps to make them more useful for statistical purposes. In this section we will deal with the two primary issues of obtaining access to administrative records, and influencing their content, design or associated procedures.

6.1 Access

The legal authority for access to administrative records is provided by Section 12 of the Statistics Act (1971):

"A person having the custody or charge of any documents or records that are maintained in any department or in any municipal office, corporation, business or organization, from which information sought in respect of the objects of this Act can be obtained or that would aid in the completion or correction thereof, shall grant access thereto for those purposes to a person authorized by the Chief Statistician to obtain such information or such aid in the completion or correction of such information."

While this provision appears to give fairly broad access rights, it is not without limitations. In some cases, legislation governing the administrative process places restrictions on access or secondary use of the administrative data. This leads to a confrontation of legislation that will at best delay the negotiation of access. In some cases, access for statistical purposes is specifically permitted.

Enabling legislation is a necessary but not sufficient condition for the productive utilization of administrative records. A co-operative approach to the development and utilization of administrative records for statistical purposes is likely to be far more effective in obtaining access to administrative records than an approach involving legal arguments and sanctions. Indeed, once access is obtained, the subsequent step of influencing design or procedures is only achievable if there is a spirit of co-operation between the administrative and statistical agencies.

Access to administrative records by Statistics Canada is strictly a one-way street. Individual micro-data are provided from the administrative agency to the statistical agency, but only confidentiality-protected aggregate data can flow back. The only exception to this rule is the case where the administrative agency depends on the statistical agency to organize, format, edit, process, or restructure its records, and a version of the original micro-data is passed back to the supplying agency.

6.2 Influencing Change

We have already alluded to the potential impact of changes in administrative regulations or practices on resulting statistics. Discontinuities in time series based on administrative records can be caused by simple changes in the coverage of a program, the introduction of an incentive to join or leave a program, or procedural changes that affect quality or completeness of records. Thus the statistical agency has to guard against, and react to, externally imposed changes.

There are other kinds of changes that the statistical agency might like to see implemented. A frequent frustration of the statistician trying to use administrative records is the feeling that the administrative records could be so much more useful if only relatively minor changes were made. For example, the addition of an extra question, the use of a different concept, the coverage of an additional subgroup, or the introduction of a quality check might significantly enhance the statistical value of the records. On the other hand, why should the administrative agency contemplate changes not required for the primary administrative process, changes which would probably in some measure add to the cost and complexity of the administrative process?

The challenge for a statistical agency is to persuade the administrators that the benefits from such a change outweigh any additional administrative costs. This is made harder to the extent that the benefits do not accrue to the department responsible for the administrative system, but to separate policy-making departments and other statistical users.

It is usually easier to build statistical requirements into a system from its inception than to make changes to a system that is already operational. Therefore, a mechanism that would allow statistical requirements to be considered during the design, or the major redesign, of an administrative system is preferable to one that only tries to adjust existing systems. A topical case in Canada is in the area of tax reform, currently under consideration by the government. This could significantly change the collection of business data in Canada. Involvement of statisticians in the design of such a system could greatly enhance the statistical benefits derived from the system. Of course, the institution of a new administrative system is a relatively rare occurrence, so that adjustment to existing systems is also necessary if statistical benefits

are to be obtained in the short run. On the other hand, the comparative rarity of design or redesign of major administrative systems strengthens the argument for not missing opportunities to influence such exercises when they do arise.

6.3 Mechanisms

A variety of measures or mechanisms, some bilateral involving the statistical agency and a specific administrative department, others of a broad government-wide nature, can assist the statistical agency in accessing and influencing administrative systems. These include:

- (i) bilateral committees at a senior level to review and discuss issues of mutual interest, including problems related to the supply of administrative data;
- (ii) feedback of statistical data to the administrative agency to demonstrate both usefulness of the data and, perhaps, weaknesses arising from administrative practices;
- (iii) provision of technical advice or services in support of the administrative agency's own statistical activities;
- (iv) a government information collection policy that requires, for example, any data collection activity plan (statistical or administrative) to be reviewed by a central agency;
- (v) statistical planning in the form of a requirement that each new program proposal include a plan for acquiring the statistical information needed to monitor and evaluate the program;
- (vi) promotion of the use of standard statistical definitions (e.g., family, business establishment, unemployed) in administrative systems;
- (vii) audits that identify the use of administrative records as a cost-efficient alternative to other means of acquiring information;
- (viii) political instruction to make greater use of particular administrative systems or seek alternatives to survey-taking;
- (ix) removal of legislative impediments to access or use of administrative records for statistical purposes.

Statistics Canada's experience in dealing with other federal government departments has been most successful in cases where close bilateral arrangements have been developed. The introduction of senior bilateral committees in the early 1980s was supportive of such arrangements, and in some cases instrumental in creating them. Government-wide measures such as information management and statistical planning have been less successful in facilitating administrative record use. Government audits and cabinet directives have provided impetus to activities aimed at increasing administrative data use, but the increased use itself is again dependent upon close working relationships with particular departments. While it is convenient to characterize the statistical agency as the progressive agency trying to break down unreasonable barriers to administrative data use, it must also be recognized that there may be inertia to the associated changes within the statistical agency itself. Staff whose careers have been based on survey design and survey-taking may need convincing that budgetary restrictions and data needs now necessitate combining these with other approaches.

Since the above comments have focused on federally administered systems, we will add a few words about provincial records. While some of the above measures apply equally to provincially administered records, the fundamental problem in dealing with subnational

jurisdictions is that of adherence to common standards. Differing provincial needs and priorities, facilitated by increasing technological capacity, will lead to divergent administrative systems in the absence of any centralizing force. Statistics Canada has used a variety of mechanisms in the past in attempts to encourage conformity, but with only mixed success. As with federal government custodians of administrative records, mutual benefit has to be the major incentive to conformity. Federal-provincial committees exist in several subject areas. The Vital Statistics Council, consisting of provincial registrars of vital events and representatives of Statistics Canada, is a successful and long-standing example. Such committees have developed and monitored conventions for reporting certain data items in the past. For example, the framework for municipal finance reporting was developed as a result of federal-provincial meetings on municipal financial statistics.

7. CONFIDENTIALITY, PRIVACY AND PUBLIC RELATIONS ISSUES

Even with the legal authority to exploit administrative records and co-operative administrative agencies to supply them, careful consideration has to be given to the public perception of the use of administrative records beyond their original purpose. Since the effectiveness, if not the survival, of a statistical agency depends critically upon the continuing co-operation and trust of respondents, it must take extreme care before embarking on any activity with the potential to undermine that co-operation or trust.

Public awareness and concern over privacy and related issues of information access and control have risen in many countries in recent years. In Canada, passage of the Privacy Act in 1982 bore witness to this mounting concern. The Privacy Act requires, *inter alia* and with some exceptions, that an index of all personal information banks under the control of federal government institutions be published periodically, that individuals have the right of access to information about themselves contained in such information banks, and that personal information be used only for purposes consistent with the purpose for which it was obtained. One of the exceptions to this last provision is that personal information may be disclosed

“... to any person or body for research or statistical purposes if ... the purpose for which the information is disclosed cannot reasonably be accomplished unless the information is provided in a form that would identify the individual to whom it relates, and ... a written undertaking (is obtained) that no subsequent disclosure of the information will be made in a form that could reasonably be expected to identify the individual to whom it relates.” (Privacy Act 1982 Section 8(2)(j)).

This provision covers the use of administrative records for statistical purposes as far as the Privacy Act is concerned. However, this Section is subject to any other Act of Parliament so that a clause forbidding such use in an Act governing an administrative process would have precedence.

While the Privacy Act and other Acts recognize statistical work as a legitimate secondary use of administrative records under certain conditions, this alone will not allay public concern over the existence of data banks that could be used to an individual's detriment. It is doubtful whether the average citizen appreciates the distinction between statistical use, where the identity of the individual record is of no lasting interest, and administrative use, where the essence of the individual record is the particular unit to which it relates. It would be easier to explain and utilize this distinction if we could state unequivocally that identifiers are never needed for statistical purposes. Unfortunately this is not the case. Several legitimate statistical

techniques do require identifiers in intermediate data manipulations. These techniques all involve some form of matching data from different files or different occasions, and identification is required to ensure that the correct records are matched. Once the matching has been accomplished the records can be anonymized provided no subsequent linkage is planned. Examples include the requirement for names in a population census to ensure coverage and permit coverage measurement, longitudinal studies using administrative records, epidemiological investigations, and evaluation studies to check survey responses against administrative sources. Explaining why identifiers are needed when identity is of no interest is an interesting challenge facing the statistical agency.

A further source of concern may relate to the undertaking of confidentiality itself. Despite Statistics Canada's record of confidentiality protection there are doubtless respondents who are skeptical about the protection their information enjoys. This concern may be heightened by the use of enumerators who are known to respondents, particularly in small communities. Some respondents seem to assume there is a high degree of information exchange actually taking place between federal departments, and in some cases do not distinguish between different departments of government.

An additional concern may relate, not to the trustworthiness of the present custodians of information banks, but to a fear that personal information cannot be protected against future violation, either illegally, or by a legitimate elected authority with different views on privacy. Protection against this possibility would require the removal of all identifying information from statistical data bases.

This public concern over privacy and the manipulation of personal information requires the statistical agency to consider measures it can take to prevent or minimize negative public reaction to its legitimate use of administrative records for statistical purposes. Since this is essentially an issue of public perception, it is important that the statistical agency be open about its practices, and that any of the following measures that are implemented are clearly visible to the interested public.

- (a) Public communications to respondents and users should continually stress the importance attached to confidentiality of all individual (micro) data acquired by the statistical agency.
- (b) The one-way nature of micro-data flow should be stressed. Micro-data flow into the statistical agency, but only confidentiality-protected aggregates or summaries flow out. This applies equally to survey or census data and data from administrative records.
- (c) The benefits of administrative record use in terms of reduced respondent burden and savings to the taxpayer should be emphasized. Such claims should be supportable by real measures of cost and respondent burden savings.
- (d) An explicit and public policy on record linkage stipulating the conditions under which the statistical agency will undertake such activities can be helpful both in demonstrating careful consideration and control of linkage activities, and in forestalling linkage requests that would violate the conditions.
- (e) The Privacy Act requires that individuals be informed of the purpose for which any personal information is being collected. Administrative agencies should be encouraged to ensure that statistical purposes are included in such statements. Even though statistical purposes may be a permissible secondary use of administrative records, their explicit mention on the collection form will serve to avoid subsequent surprise.

- (f) The physical security that surrounds the use of sensitive administrative records should be clearly visible, and perhaps even tighter than that in use generally within the Agency. For example, in Statistics Canada, the divisions having primary custody of tax data are housed in limited access areas within buildings that are themselves subject to security checks on entry.
- (g) Exemption of statistical files from examination by security or intelligence services is an important element in maintaining public trust in the absolute confidentiality of data provided to the statistical agency. An exemption for Statistics Canada data (the sole institutional exception within government) was provided when the new Canadian Security and Intelligence Service was formed in 1983.

While the above points represent some specific measures that can be taken to avoid or respond to public reaction to the use of administrative records, ultimately the statistical agency must have strong political support for this kind of activity. The political credit to be gained from demonstrated reductions in costs and respondent burden, coupled with strong political assurances of the protection of individual data, provide a strong platform for politicians to dispel public concern over the use of administrative records for statistical purposes. At the same time they must immediately and unambiguously confront and correct any suggestion that statistical records be used for administrative purposes.

8. CONCLUSION

Administrative records are and will continue to be an increasingly important source of statistical data. The relative strengths and weaknesses of data derived from administrative systems, in terms of cost, coverage, quality, relevance and timeliness, in comparison to census- or survey-based data, dictate the manner in which these sources of data are most effectively used. Current uses of administrative records include direct tabulation, indirect estimation, substitution for survey responses, frame construction and maintenance, and data evaluation. These uses now permeate most statistical programs and can be expected to extend even further in the future.

In Canada, administrative records have become part of the fabric of our statistical system. Their use has been one of the means by which Statistics Canada has been able to maintain its programs in the face of declining budgets. In the process, respondent burden has been reduced and new, or more frequent, data series have become available. Since we do not have administrative registers as such, considerable attention has been paid to issues of coverage and the joint use of both administrative and survey-based data to ensure valid estimation of universe totals. The use of record linkage techniques, though requiring careful controls, has proven to be very valuable, particularly for business data, longitudinal labour market studies, and epidemiological work.

With the growing use of administrative records, statistical agencies are becoming increasingly dependent upon other agencies for the uninterrupted flow of input data to their statistical programs. Whatever the legislative and policy environment in which the statistical agency operates, the establishment of close co-operative arrangements with supplying agencies is crucial. The ability of the statistical agency to influence the design or redesign of administrative systems rests on a mutual understanding of the requirements of the two agencies. Establishment of a government-wide policy or principle that the statistical agency should have a voice in decisions regarding the design of administrative systems, or more generally, in proposals

for meeting the statistical needs of new programs, can help the statistical agency in this regard, but is no substitute for the fostering of close co-operation with administrative agencies.

A variety of mechanisms can be considered to assist the statistical agency in gaining the access and influence it requires within the government system. The applicability and effectiveness of each mechanism will depend upon the underlying legislative and political climate, and on the mandate and status of the statistical agency within the government apparatus. Statistics Canada's experience has been that close bilateral working relationships with administrative departments, based on a principle of mutual benefit, is the most effective approach. Political support for the use of administrative records is important and has been forthcoming through recent government decisions related to budget reductions.

ACKNOWLEDGEMENTS

This paper has drawn upon ideas and work of many people at Statistics Canada. In particular, parts of this paper incorporate material prepared in collaboration with Michael Colledge, Ivan Fellegi and John Leyes.

REFERENCES

- AUGER, E. (1987). Family data from the Canadian personal income tax file. Paper presented at 1987 American Statistical Association meetings.
- BRACKSTONE, G.J. (1984). The impact of technological change on census-taking, *Estadística*, 36, 43-60.
- CANADA, STATISTICS ACT (1971). *Statutes of Canada 1970-71-72*, c.15.
- CANADA, PRIVACY ACT (1982). *Statutes of Canada 1980-81-82*, c.11.
- CHOUDHRY, G.H., and HIDIROGLOU, M.A. (1987). Small area estimation: Some experiences at Statistics Canada. *Proceedings of the 46th Session of the International Statistical Institute*, (forthcoming).
- COLLEDGE, M.J. (1987). The Business Survey Redesign Project: Implementation of a new strategy at Statistics Canada. *Proceedings of the Third Annual Research Conference, U.S. Bureau of the Census*, (forthcoming).
- FEENEY, G.A. (1987). The estimation of the number of unemployed at the small area level. In *Small Area Statistics, An International Symposium*, (Eds. R. Platek, J.N.K. Rao, C.E. Sarndal and M.P. Singh), New York: John Wiley.
- NORRIS, D.A., and STANDISH, L.D. (1983). A technical report on the development of migration data from taxation records. Working Paper, Statistics Canada.
- ROWEBOTTOM, L.E. (1978). The utilization of administrative records for statistical purposes. *Survey Methodology*, 4, 1-15.
- STATISTICS CANADA (1987). *Urban FSA and rural postal code summary data*. Catalogue No. 17-602, Statistics Canada.
- TROTTIER, I., and CHOUDHRY, G.H. (1985). Model based unemployment estimates for small areas. In *Small Area Statistics, An International Symposium '85 (Contributed Papers)*, (Eds., R. Platek and M.P. Singh), Laboratory for Research in Statistics and Probability, Carleton University/University of Ottawa.

Statistical Properties of Crop Production Estimators

CAROL A. FRANCISCO, WAYNE A. FULLER, and RON FECOSO¹

ABSTRACT

The National Agricultural Statistics Service, U.S. Department of Agriculture, conducts yield surveys for a variety of field crops in the United States. While field sampling procedures for various crops differ, the same basic survey design is used for all crops. The survey design and current estimators are reviewed. Alternative estimators of yield and production and of the variance of the estimators are presented. Current estimators and alternative estimators are compared, both theoretically and in a Monte Carlo simulation.

KEY WORDS: Crop surveys; Yield estimation; Two phase sample; Variance estimation.

1. INTRODUCTION

The National Agricultural Statistics Service (formerly known as the Statistical Reporting Service), U.S. Department of Agriculture, conducts objective yield surveys of corn, cotton, soybeans, rice, grain sorghum, sunflowers and wheat in states which are major producers of these field crops. Similar yield surveys are conducted in a number of other countries.

While field sampling procedures for each crop differ in terms of plot sizes, plot location methods, and vegetative and fruit measurement techniques, all surveys rely on the same basic design. A four-step sampling procedure is used. A description of this survey design is contained in Section 2. Section 3 describes the estimators of average crop yield and the variance estimators, evaluates them and explores alternative estimators. Conclusions and recommendations are presented in Section 4.

2. OBJECTIVE YIELD SURVEY DESIGN

The first two steps of sample selection produce the sample of area segments used in the June Enumerative Survey conducted by the National Agricultural Statistics Service (NASS). The area frame for each state is stratified by land use. For example, the State of California is divided into 12 land use strata. Each land use stratum is subdivided into areas called frame units. The size of a frame unit varies; the actual size of any given frame unit depends upon available boundary designations, available ancillary information, political boundaries, and so forth. Once frame units are established, the number of area segments in each frame unit is determined by dividing the total area of each frame unit by the target segment size. The target size is a function of the land use stratum into which the frame unit falls. For example, in California the target segment size is one half square mile in the orchard stratum and one square mile in all other cropland strata. Frame units typically contain between one and 30 area segments.

¹ Carol A. Francisco, Syntex Laboratories Inc., 3401 Hillview Avenue, Palo Alto, California 94304; Wayne Fuller, Department of Statistics, Iowa State University, Ames, Iowa 50011; and Ron Fecso, Survey Research Branch, National Agricultural Statistics Service, U.S. Department of Agriculture, Washington D.C. 20250.

Each land use stratum is substratified on the basis of geography. To develop the geographic substrata, frame units within each land use stratum are ordered by county in such a manner that adjacent counties that are agriculturally similar are placed together (Fecso 1978). Substrata are formed from sequential groups of area segments. Thus, substrata contain area segments that are agriculturally similar and geographically close together. Within a given land use stratum, substrata have an equal number of segments and equal area (within rounding). Detailed information on the area frame design is available in Fecso and Johnson (1981) and Houseman (1975).

For purposes of variance estimation, it is the substrata within land use strata that are the sampling strata. Henceforth, the land use substrata will be referred to simply as strata.

The first step in sampling from the area frame is the selection of frame units within each stratum. The number of frame units allocated to a stratum depends on the agricultural nature of the stratum. Typically, eight to 15 frame units are drawn in cropland strata; whereas in agri-urban, city, and nonagricultural strata four to five frame units are drawn. Frame units within strata are selected at random with probability proportional to the number of area segments in the frame unit. At the second step, one area segment is chosen at random from each selected frame unit. Thus, each area segment within a stratum has an equal probability of selection.

Although the frame unit is the primary sampling unit for this design, because the frame units are selected with probability proportional to the number of segments and one segment is selected per sampled frame unit, the segment can be treated as the primary sampling unit. In our study, steps one and two in the sampling procedure are considered as one procedure, and the sample of segments will be treated as a stratified single stage simple random sample. Since the average sampling rate is about one percent, the finite population correction term will be ignored in our analysis.

The third and fourth steps in the sampling procedure involve the selection of fields and of plots within selected fields. As part of the June Enumerative Survey, all selected area segments are screened for fields which have been planted or are scheduled to be planted with the crop of interest. These fields are listed by segment number and order of enumeration within segment. A systematic sample of fields is selected with selection probabilities proportional to the product of the field area and the inverse of the probability of selection of the area segment in which the field is contained. Hence, the number of sampled fields per segment varies, and large fields within a segment can be selected more than once.

At the fourth and final step, two plots of roughly equal area are placed in each selected field using a random row and pace method of location. Where rows are not readily distinguishable, and in the case of wheat, a random number of paces along the field edge and a random number of paces into the field are used to locate plots. A further exception occurs in the wheat objective yield survey. For this survey the first plot is randomly located and the second plot is placed in a fixed position relative to the first plot. In the event that a large field is selected more than once during the third step of the sampling procedure, additional sets of two plots are independently sampled. Because plots are always sampled in pairs, we call the pair of plots the secondary unit. A maximum of eight plots (that is, four secondary units) per field is imposed.

3. ESTIMATION PROCEDURES

Formally, the sample is a two phase sample with subsampling in the second phase. Table 1 contains a schematic description of the sample. The phase one sample is a stratified simple

Table 1
Sampling Procedure for the Objective Yield Survey

Phase/Sampling Unit	Selection Procedure	Sampled Number ¹	Data Collected
Phase One			
Primary Sampling Unit: Segment	equal probability within strata	n_h	crop acres
Phase Two			
Primary Sampling Unit: Segment	unequal probability	K_h	crop acres, estimated production ²
Secondary Sampling Unit: Pair of Plots	equal probability	m_{hk}	estimated production from plots

¹ Number is per stratum for primary sampling units and is per segment for secondary sampling units.
² Segment production is zero if the crop acreage is zero and is estimated from plot determinations if the crop acreage is positive.

random sample of segments. The phase two sample is composed of all segments with zero crop acres and a probability-proportional-to-crop-acres sample of segments with the crop. The sample of segments is the result of a probability-proportional-to-area systematic sample of first phase fields planted with the crop. A sample of secondary units, where each secondary unit is a pair of plots, is selected from the segments in the phase two sample that have the crop. Because the secondary unit is always a pair of plots, we will henceforth refer to secondary units and no longer speak of plots. We will also ignore the fact that the operational units used to locate the plots are fields and speak only of the sampled segments.

Notice that two types of segments are observed at phase two – those that have zero acres of the crop and those that have non-zero acres. The total number of second phase segments is K . The acres and the total production are known (both equal to zero) for an observed segment with zero acres. For second phase segments with positive acres, a subsample of secondary units is used to estimate production.

Let M_{hk} be the number of secondary units in segment k of the h -th stratum. Without loss of generality, M_{hk} could be assumed to be equal to A_{hk} , where A_{hk} is the crop area in segment hk . Equality requires only the choice of an appropriate scale for area.

Section 3.1 examines the yield estimator that is currently used. Conditions under which this estimator is unbiased for state average yield are investigated. A simple estimator of the variance of estimated yield is discussed in Section 3.2. Estimators of the unconditional variances of the yield and production estimators are developed in Section 3.3. A Monte Carlo study of estimators is given in Section 3.4.

3.1 Currently Used Yield and Production Estimators

Estimates of the state average yield are currently computed as though the sample were an equal probability simple random sample of secondary units. The estimator is the simple

average yield of secondary units with positive acreages. That is, the estimated average yield per acre is

$$\bar{y} = D^{-1} \sum_{h=1}^L \sum_{k=1}^{n_h} \sum_{\ell=1}^{m_{hk}} Y_{hk\ell} \delta_{hk\ell}, \quad (3.1)$$

where

$$\delta_{hk\ell} = 1 \quad \text{if } A_{hk} > 0,$$

$$\delta_{hk\ell} = 0 \quad \text{if } A_{hk} = 0,$$

$$D = \sum_{h=1}^L \sum_{k=1}^{n_h} \sum_{\ell=1}^{m_{hk}} \delta_{hk\ell}, \quad (3.2)$$

m_{hk} is the number of sampled secondary units selected in segment hk , L is the number of strata, and $Y_{hk\ell}$ is the estimated yield per acre for secondary unit ℓ of segment hk . If the crop acreage in a segment, A_{hk} , is zero, then $m_{hk} = 1$ and $Y_{hk\ell} = 0$, by definition. The total number of observed secondary units for segments with positive acres is D .

Expression (3.1) can be written in the convenient operational form

$$\bar{y} = D^{-1} \sum_{t=1}^D Y_t, \quad (3.3)$$

where the subscript t replaces the triple subscript $hk\ell$ and the summation is over secondary units in segments with positive crop acres.

The estimator of average crop yield per acre (3.1) is a type of combined ratio estimator. This can be shown by using conditional selection probabilities to rewrite \bar{y} . In the NASS scheme, segments are selected systematically with probabilities proportional to expanded size, and segments with sufficiently large expanded acreage are included with certainty. The number of secondary units allocated to certainty segments is proportional to the size of the segment, up to rounding error. The rounding is performed by the systematic selection scheme. Let $\pi_{hk\ell}$ be the conditional probability that secondary unit ℓ in segment k of stratum h is selected, given the sample of segments selected at the first phase of the sampling procedure. We have

$$\pi_{hk\ell} = D \left(\sum_{h=1}^L N_h n_h^{-1} \sum_{k=1}^{n_h} M_{hk} \right)^{-1} N_h n_h^{-1} \quad (3.4)$$

for secondary units in segments with $A_{hk} > 0$, where N_h is the population number of segments in stratum h , M_{hk} is the number of secondary units in segment k of stratum h , and n_h is the number of segments in stratum h selected at the first phase. The conditional probability of observing a segment with zero acres at the second phase is one.

Then the mean estimator given in (3.1) can be written as

$$\bar{y} = \frac{\sum_{h=1}^L N_h n_h^{-1} \sum_{k=1}^{K_h} \sum_{\ell=1}^{m_{hk}} \pi_{hk\ell}^{-1} Y_{hk\ell}}{\sum_{h=1}^L N_h n_h^{-1} \sum_{k=1}^{K_h} \sum_{\ell=1}^{m_{hk}} \pi_{hk\ell}^{-1} \delta_{hk\ell}}, \quad (3.5)$$

where $N_h n_h^{-1}$ is the inverse of the first stage selection probability, K_h is the number of second phase segments drawn from stratum h , and $K = \sum K_h$. Given an appropriate scale, the numerator of (3.5) is an estimator of the total production and the denominator is an estimator of the total area. It can be shown that the numerator is an unbiased estimator by taking expectations, conditioning on the first phase sample units and then averaging over first phase samples. The denominator is a stratified estimator of the total number of secondary units. By the nature of the sampling, the number of sampling units is proportional to acreage and one can choose the scale so that the number of secondary units is equal to acreage. Hence, \bar{y} can be viewed as the ratio of an unbiased estimator of the total production of the crop to an unbiased estimator of the total area under the crop.

To estimate total state production, NASS multiplies \bar{y} by \hat{A} , where \hat{A} is the estimator of total crop acreage defined by

$$\hat{A} = \sum_{h=1}^L N_h n_h^{-1} \sum_{k=1}^{n_h} A_{hk}. \quad (3.6)$$

Thus, the estimated total production is

$$\hat{Y} = \hat{A} \bar{y}. \quad (3.7)$$

3.2 Simple Variance Estimators

Under the assumption of simple random sampling of secondary units from the entire set of secondary units available at the second phase, the estimated variance of \bar{y} conditional on the second phase segments is

$$\hat{V}_2(\bar{y}) = D^{-1} (D - 1)^{-1} \sum_{t=1}^D (Y_t - \bar{y})^2, \quad (3.8)$$

where the subscript 2 on \hat{V} is used to denote conditional variance and the subscript t on Y replaces the triple subscript $hk\ell$. The sum over t is the sum over the D secondary units in segments with positive acres.

Because of the simplicity of expression (3.8), it has been suggested that it be used as an estimator of the unconditional variance. It has also been suggested that the variance of the estimated total state production be estimated with

$$\hat{V}_*(\hat{Y}) = \hat{A}^2 \hat{V}_2(\bar{y}) + \bar{y}^2 \hat{V}(\hat{A}) + \hat{V}(\hat{A}) \hat{V}_2(\bar{y}), \quad (3.9)$$

where \hat{A} is defined in (3.6) and $\hat{V}(\hat{A})$ is the usual variance estimator for a stratified estimated total,

$$\hat{V}(\hat{A}) = \sum_{h=1}^L N_h^2 n_h^{-1} (n_h - 1)^{-1} \sum_{k=1}^{n_h} (A_{hk} - \bar{A}_h)^2, \quad (3.10)$$

and

$$\bar{A}_h = n_h^{-1} \sum_{k=1}^{n_h} A_{hk}.$$

The estimator (3.9) is an estimator of the variance of a product based on an implicit assumption that \bar{y} and \hat{A} are uncorrelated.

Evaluation of the extent to which the estimator (3.9) tends to underestimate the variance of \hat{Y} is difficult. We can express the unconditional variance of \bar{y} as

$$\begin{aligned} V(\bar{y}) &= V_1 \{E_2(\bar{y})\} + E_1 \{V_2(\bar{y})\} \\ &= V_1 \{\hat{A}^{-1} \sum_{h=1}^L N_h n_h^{-1} \sum_{k=1}^{n_h} Y_{hk}\} + E_1 \{V_2(\bar{y})\}, \end{aligned} \quad (3.11)$$

where $Y_{hk} = M_{hk} \bar{Y}_{hk}$ is the total for the k -th segment in stratum h , and E_1 and V_1 denote the expectation and variance, respectively, with respect to first phase sampling.

The estimator $\hat{V}_2(\bar{y})$ is unbiased for the second component of expression (3.11) under simple random sampling of secondary units. Because sampling at phase two of the NASS scheme is done systematically, $\hat{V}_2(\bar{y})$ is a biased estimator of $V_2(\bar{y})$. The nature and extent of this bias depends upon the correlation structure of the list used in sample selection at the second phase. Also affecting the bias in $\hat{V}_2(\bar{y})$ as an estimator of the true variance is the fact that formula (3.8) was derived under an assumption of replacement sampling at phase two. To the extent that phase two sampling is actually done without replacement (because samples are drawn systematically from the list of expanded segment acreages, a segment is sampled more than once only if it is large), $\hat{V}_2(\bar{y})$ will overestimate $V_2(\bar{y})$.

The estimator $\hat{V}_*(\hat{Y})$ contains no estimator of $A^2 V_1 \{E_2(\bar{y})\}$, and this produces a negative bias. However, estimation of that component is not easy, even under the simplifying assumption of probability-proportional-to-size sampling at phase two. Because of these considerations, the performance of $\hat{V}_*(\hat{Y})$ will be studied by Monte Carlo methods in Section 3.4.

3.3 Alternative Estimators of Variance

An alternative approach to the estimation of $V(\bar{y})$ is to view the sample as a two phase sample, as shown in Table 1, and to assume that the unconditional probability of selecting a segment to receive a secondary unit is proportional to the conditional probability given the first phase segments.

Let π_{hk} be the conditional probability that segment k in stratum h is included in the second phase, given the first phase sample of segments. We have

$$\pi_{hk} = \min(1, M_{hk} \pi_{hkt}), \quad (3.12)$$

where $\pi_{hk\ell}$ is a constant within segment hk . If $\pi_{hk} = 1$ and the segment is selected to receive more than one secondary unit, it is assumed that the secondary units are independently drawn.

Let π_{hk}^* be the unconditional probability that an observation is made on segment k in stratum h at phase two. If $A_{hk} = 0$, then π_{hk}^* is the unconditional probability that segment hk is selected to receive at least one secondary unit. If $A_{hk} = 0$, then π_{hk}^* is equal to the probability that segment hk is selected at the first phase of sampling. Let

$$\begin{aligned} \pi_{hk}^* &= \frac{n_h}{N_h} && \text{if } A_{hk} = 0, \\ \pi_{hk}^* &= \pi_{hk} \frac{n_h}{N_h} && \text{if } 0 < \pi_{hk} < 1, \end{aligned} \tag{3.13}$$

where π_{hk} , defined in (3.12), is the conditional probability that the hk -th segment is selected in phase two, given the first phase sample.

In our analysis we assume the π_{hk}^* to be fixed. This will be so and the probability π_{hk}^* will be the true unconditional probability if π_{hk} is a specified multiple of M_{hk} where the multiple is fixed before sample selection. Expression (3.13) will be an approximation if π_{hk} is a function of the segments selected at the first step of the selection procedure.

Expression (3.13) is proportional to M_{hk} for $M_{hk}\pi_{hk} \leq 1$. If $M_{hk}\pi_{hk\ell} > 1$, then the number of selected secondary units is greater than or equal to one. The correct number of secondary units to allocate to such segments to maintain a self-weighting sample of secondary units is $M_{hk}\pi_{hk\ell}$. In practice, the number of secondary units observed as a result of probability-proportional-to-size systematic sampling never differs from $M_{hk}\pi_{hk\ell}$ by more than one.

To simplify the remaining computations, we assume that the systematic sampling design contains no rounding error. In other words it is assumed that the number of secondary units observed per segment is equal to the number required for a self-weighting sample. Thus, it is assumed that the number of secondary units observed in a segment drawn as part of the second phase of sampling is

$$\begin{aligned} m_{hk} &= 1 && \text{if } 0 < \pi_{hk} < 1, \\ m_{hk} &= M_{hk}\pi_{hk\ell} && \text{if } \pi_{hk} = 1. \end{aligned} \tag{3.14}$$

Under this assumption, an unequal probability combined ratio estimator of the mean yield is equivalent to estimator (3.1). The combined ratio estimator is

$$\bar{y}_r = \hat{M}_r^{-1} \sum_{h=1}^L \sum_{k=1}^{K_h} \pi_{hk}^{*-1} M_{hk} \bar{y}_{hk}, \tag{3.15}$$

where

$$\begin{aligned} \bar{y}_{hk} &= m_{hk}^{-1} \sum_{\ell=1}^{m_{hk}} Y_{hk\ell} && \text{if } A_{hk} > 0, \\ \bar{y}_{hk} &= 0 && \text{if } A_{hk} = 0, \end{aligned}$$

$$\hat{M}_r = \sum_{h=1}^L \sum_{k=1}^{K_h} \pi_{hk}^{*-1} M_{hk}.$$

In expression (3.15) and the remaining expressions of this section, the reader can read \hat{A}_r (total area) for \hat{M}_r (total secondary units), if so desired.

In the following discussion, replacement sampling of segments with probabilities proportional to the area of a crop within the segment is assumed as an approximation to the probability-proportional-to-size systematic sampling scheme of the second phase. An estimator of the variance of \bar{y} under the assumption of replacement sampling is

$$\hat{V}(\bar{y}_r) = \hat{M}_r^{-2} \sum_{h=1}^L K_h (K_h - 1)^{-1} \sum_{k=1}^{K_h} (\pi_{hk}^{*-1} u_{hk} - \bar{u}_h)^2, \quad (3.16)$$

where

$$u_{hk} = M_{hk} (\bar{y}_{hk} - \bar{y}_r),$$

$$\bar{u}_h = K_h^{-1} \sum_{k=1}^{K_h} \pi_{hk}^{*-1} u_{hk}.$$

An estimator of the total production is

$$\hat{Y}_r = N \bar{M}_n \bar{y}_r, \quad (3.17)$$

where

$$\bar{M}_n = \sum_{h=1}^L W_h n_h^{-1} \sum_{k=1}^{n_h} M_{hk}$$

N is the total number of segments in the population and $W_h = N^{-1} N_h$. The Taylor approximation of the unconditional variance of the approximate distribution of \hat{Y}_r is

$$V\{\hat{Y}_r\} = N^2 [\bar{M}_N^2 V\{\bar{y}_r\} + 2\bar{M}_N \bar{y}_N C\{\bar{y}_r, \bar{M}_n\} + \bar{y}_N^2 V\{\bar{M}_n\}], \quad (3.18)$$

where \bar{y}_r is given in (3.15), \bar{M}_n is defined in (3.17),

$$\bar{M}_N = N^{-1} \sum_{h=1}^L \sum_{k=1}^{N_h} M_{hk},$$

$$\bar{y}_N = \left(\sum_{h=1}^L \sum_{k=1}^{N_h} M_{hk} \right)^{-1} \sum_{h=1}^L \sum_{k=1}^{N_h} Y_{hk},$$

$Y_{hk} = M_{hk} \bar{y}_{hk}$ is the total for the k -th segment in stratum h , and $C\{\bar{y}_r, \bar{M}_n\}$ is the covariance between \bar{y}_r and \bar{M}_n .

Under the unequal-probability-fixed-take procedure, the estimator $\bar{y}_r (\doteq \bar{y})$ is approximately conditionally unbiased for the mean yield for the $n = \sum n_h$ segments in the first phase sample. The mean yield of the n segments is

$$\bar{y}_n = \bar{M}_n^{-1} \sum_{h=1}^L W_h n_h^{-1} \sum_{k=1}^{n_h} Y_{hk}.$$

Therefore, the covariance between \bar{y}_r and \bar{M}_n is the covariance between $\bar{M}_n^{-1} \bar{Y}_n$ and \bar{M}_n , where

$$\bar{Y}_n = \sum_{h=1}^L W_h n_h^{-1} \sum_{k=1}^{n_h} Y_{hk}.$$

Using the common approximation for a ratio, the covariance between \bar{y}_r and \bar{M}_n can be approximated by

$$\begin{aligned} C\{\bar{M}_n^{-1} \bar{Y}_n, \bar{M}_n\} &\doteq C\{(\bar{Y}_n - \bar{y}_N \bar{M}_n) \bar{M}_n^{-1}, \bar{M}_n\} \\ &= \bar{M}_n^{-1} [C\{\bar{Y}_n, \bar{M}_n\} - \bar{y}_N V\{\bar{M}_n\}]. \end{aligned} \quad (3.19)$$

If the probability of observing the pair (Y_{hk}, M_{hk}) is proportional to π_{hk}^* , an estimator of the covariance between \bar{Y}_n and \bar{M}_n is

$$\hat{C}\{\bar{Y}_n, \bar{M}_n\} = \sum_{h=1}^L W_h^2 n_h^{-1} \hat{S}_{MYh} \quad (3.20)$$

where

$$\hat{S}_{MYh} = K_h (K_n^{-1})^{-1} \left(\sum_{j=1}^{K_h} \pi_{hj}^{*-1} \right)^{-1} \sum_{k=1}^{K_h} \pi_{hk}^{*-1} (M_{hk} - \bar{M}_h^*) (M_{hk} \bar{y}_{hk} - \bar{y}_{h..}^*),$$

$$\bar{M}_h^* = \left(\sum_{j=1}^{K_h} \pi_{hj}^{*-1} \right)^{-1} \sum_{k=1}^{K_h} \pi_{hk}^{*-1} M_{hk},$$

$$\bar{y}_{h..}^* = \left(\sum_{j=1}^{K_h} \pi_{hj}^{*-1} \right)^{-1} \sum_{k=1}^{K_h} \pi_{hk}^{*-1} M_{hk} \bar{y}_{hk}.$$

The estimator \hat{S}_{MYh} is constructed as a degrees-of-freedom adjustment to a Horvitz-Thompson ratio estimator of the mean of the products $(M_{hk} - \bar{M}_h)(Y_{hk} - \bar{Y}_{h..})$. The degrees-of-freedom adjustment, the factor $K_h(K_h - 1)^{-1}$, is introduced because it is necessary to replace the population means with sample means when constructing the product.

Substituting (3.15), (3.16), and (3.20) into (3.18) gives

$$\hat{V}\{\bar{Y}_r\} = N^2 [\bar{M}_n^2 \hat{V}\{\bar{Y}_r\} + 2\bar{y}_r \hat{C}\{\bar{Y}_n, \bar{M}_n\} - \bar{y}_r^2 \hat{V}\{\bar{M}_n\}], \quad (3.21)$$

where $\hat{V}\{\bar{M}_n\}$ is the variance estimator for a stratified mean. Equation (3.21) is a stratified double sampling estimator of the variance of the estimated total state production. Unlike the estimator $\hat{V}_*(\hat{Y})$ of (3.9), estimator (3.21) does not assume that the yield and acreage estimators are uncorrelated. Equation (3.21) also uses an unconditional estimator of the variance of yield.

3.4. A Monte Carlo Comparison of Estimators

A Monte Carlo study was performed to illustrate the differences among alternative estimators. Cotton acreage data from the 1983 June Enumerative Survey in California and data from the corresponding 1983 objective yield survey were used as a basis for the study. For purposes of the Monte Carlo study, 28 strata were considered to have cotton.

Table 2 shows the distribution of cotton among the 28 strata as observed in the 1983 June Enumerative Survey. Fecso and Johnson (1981) describe the six different land uses, where land use is the first two digits of the stratum identification, as follows:

- 1300 – 50% or more cultivated land, primarily general crops with less than or equal to 10% fruit or vegetables;
- 1700 – 50% or more cultivated land, primarily fruit, tree nuts, or grapes mixed with general crops;
- 1900 – 50% or more cultivated land, primarily vegetables mixed with general crops;
- 2000 – 15-50% cultivated land with extensive cropland and hay;
- 3100 – residential mixed with agricultural lands, more than 20 dwellings per square mile;
- 4100 – less than 15% cultivated land, primarily privately owned rangeland.

A population was simulated from the results of the 1983 June Enumerative Survey. Table 2 compares the characteristics of the simulated population to the results of the survey. In the simulated population, cotton was determined to be present in segment k ($k = 1, \dots, N_h$) within stratum h ($h = 1, \dots, 28$) if $X_{hk} = 1$, where X_{hk} is an independent Bernoulli (p_h) random variable and p_h is the observed proportion of segments in stratum h found to have cotton in the 1983 June Enumerative Survey.

The next step in the creation of the population was the assignment of cotton acres to the segments for which $X_{hk} = 1$. A set of 1983 observed ratios of segment cotton acreages to the average segment acreage was compiled for land use substrata having more than one segment with cotton in the 1983 June Enumerative Survey. This set of observed ratios was used to generate the number of cotton acres in segments having cotton. If $X_{hk} = 1$, then a ratio, r_{hk} , was drawn from the set of observed ratios such that each observed ratio in the set had an equal probability of selection. The number of acres of cotton in segment hk , M_{hk} , was defined by

$$M_{hk} = r_{hk}\bar{M}_h, \quad (4.1)$$

where \bar{M}_h was the observed average number of cotton acres for segments with cotton in stratum h in the 1983 June Enumerative Survey. (See Table 2.)

Results of the 1983 objective yield survey for cotton were used to simulate yield observations within segments. Since estimated yields were not readily accessible, an alternative variable – a major component of yield estimates – was used. This variable is the number of plants per 100 square feet. The estimated overall population mean number of plants per 100 square feet was 79.6 for the 1983 objective yield survey. Table 3 shows the average number of plants

Table 2
Cotton Acreage Estimates from the 1983 June Enumerative Survey
in California and Cotton Acreages in the Simulated Population

Stratum	Target Segment Size (Acres)	Number of Segments in Stratum	Number of Segments Sampled in 1983	Percentage of Segments with Cotton		Mean Acres Cotton in Segments with Cotton	
				1983	Simulated Population	1983	Simulated Population
1314	640	291	10	60	60	197	200
1315	640	291	10	100	100	354	348
1316	640	291	10	90	89	167	173
1317	640	291	10	90	92	149	148
1318	640	291	10	50	53	481	422
1319	640	291	10	20	19	249 ¹	260
1320	640	291	10	90	91	154	155
1321	640	291	10	60	61	270	274
1322	640	291	10	70	71	205	210
1323	640	291	10	80	79	288	279
1713	320	432	10	30	28	125	122
1714	320	432	10	30	31	58	57
1715	320	432	10	20	22	86 ²	84
1716	320	432	10	10	8	86 ²	89
1717	320	432	10	40	38	26	27
1718	320	432	10	30	29	144	144
1719	320	432	10	30	31	65	67
1720	320	432	10	30	30	38	35
1721	320	432	10	30	29	133	138
1722	320	432	10	50	47	130	131
1723	320	432	10	40	40	76	76
1906	640	362	10	70	73	117	127
1907	640	362	10	70	74	192	194
1908	640	362	10	80	83	253	246
2010	640	649	10	30	31	303	306
2011	640	649	10	40	41	175	165
3107	160	1,847	5	20	22	25 ³	25
4110	2,560	1,044	10	10	10	178	165

¹ Number of segments sampled was less than or equal to 2. Average of all segments in substrata within land use stratum 13 is shown.

² Number of segments sampled was less than or equal to 2. Average of all segments in substrata within land use stratum 17 is shown.

³ Number of segments sampled was less than or equal to 2. Approximate acreages for this agri-urban stratum are shown.

per 100 square feet. The estimated overall population mean number of plants per 100 square feet was 79.6 for the 1983 objective yield survey. Table 3 shows the average number of plants per 100 square feet by stratum for the 1983 survey. The average for each stratum is based on all secondary units within the stratum that were drawn as part of the probability-proportional-to-estimated-size sampling scheme.

An analysis of variance of the 1983 plant data (Table 4) shows that 28 percent of the total variation among secondary units was due to between-segment differences within strata ($s_b^2 = 378.0$), whereas 58 percent of the total variation was due to variation among secondary units within segments ($s_w^2 = 776.6$). If the stratum component is treated as fixed, 67 percent of the within-segment variation is due to variance among secondary units.

Table 3
Average Number of Plants per 100 Square Feet from the 1983
Objective Yield Survey for Cotton in California and in the
Simulated Population

Stratum	Average Number of Plants per 100 Square Feet	
	1983 Objective Yield Survey	Simulated Population
1314	78	76
1315	80	80
1316	67	68
1317	72	73
1318	80	80
1319	93	93
1320	92	91
1321	70	69
1322	84	84
1323	72	71
1713	118	117
1714	96 ¹	95
1715	96 ¹	93
1716	96 ¹	86
1717	96 ¹	96
1718	139	140
1719	96 ¹	97
1720	96 ¹	97
1721	89	86
1722	79	79
1723	84	85
1906	98	98
1907	67	67
1908	53	53
2010	118	118
2011	47	47
3107	80 ²	79
4110	60	59

¹ Number secondary units observed was less than or equal to 2. Secondary unit average for land use stratum 17 is shown.

² Number secondary units observed was less than or equal to 2. Secondary unit average for all strata is shown.

Table 4
Analysis of Variance for the 1983 Objective Yield Survey Data

Source	Degrees of Freedom	Sum of Squares	Mean Square	Variance Component	Percent of total
Stratum	26	80,193	3,084.3	187.3	14
Segment within Stratum	85	124,086	1,459.8	378.0	28
Residual	103	79,991	776.6	776.6	58
Total	214	284,270		1,341.9	100

When a segment had cotton, the mean number of plants per 100 square feet for segment hk was simulated by

$$\bar{c}_{hk} = \bar{c}_h + e_{hk}, \tag{4.2}$$

where \bar{c}_h is the average number of plants per 100 square feet for stratum h , e_{hk} is distributed $N(0, s_b^2)$, and $s_b^2 = 378.0$. In the event that the simulated segment mean (\bar{c}_{hk}) was less than 10% of the stratum mean, then c_{hk} was set equal to $(.10)\bar{c}_h$. Table 3 compares the simulated stratum means with those from the 1983 objective yield survey. The overall mean in the simulated population was $\bar{y}_N = 79.6$.

From the simulated population 500 June Enumerative Survey samples were drawn using stratified random sampling. A total of 275 segments were drawn for each of the simulated samples. The number of segments drawn from each stratum was the same as that for the 1983 June Enumerative Survey (see Table 2). For each of the simulated samples, estimates of the mean number of acres per segment in the population, as well as the conditional probabilities π_{hk} , from (3.12), that the segments in the sample would receive plots in a draw, were calculated. These conditional probabilities were used at the second stage of sampling in the single start probability-proportional-to-estimated-size systematic sampling described in Section 2. Objective yield survey samples were simulated by selecting 220 secondary units using this systematic sampling scheme. Two objective yield survey samples were simulated for each of the 500 simulated June Enumerative Survey samples.

When a segment was selected to receive a secondary unit, the yield (number of plants per 100 square feet) observed within a field was simulated under the assumption that the coefficient of variation within each segment was constant. The observed number of plants was defined as

$$y_{hk\ell} = \bar{c}_{hk} + s_w \bar{y}_N^{-1} \bar{c}_{hk} f_{hk\ell}, \tag{4.3}$$

where $y_{hk\ell}$ is the estimated average number of plants per 100 square feet for the ℓ -th secondary unit in segment k of stratum h , and $f_{hk\ell}$ is distributed $N(0, 1)$. The within-segment standard error is the square root of the $s_w^2 = 776.6$ of Table 4, and \bar{y}_N is the overall mean number of plants per plot. In the event that $y_{hk\ell}$ was less than 10% of the stratum mean, then $y_{hk\ell}$ was set equal to $(.10)\bar{c}_{hk}$. Similarly, if $y_{hk\ell}$ was greater than 190% of the stratum mean, then $y_{hk\ell}$ was set equal to $(1.9)\bar{c}_{hk}$.

Results of the simulations for cotton acreages are summarized in Table 5. The estimated mean acres per segment is

$$\bar{A}_n = \sum_{h=1}^L W_h n_h^{-1} \sum_{k=1}^{n_h} A_{hk}, \tag{4.4}$$

with estimated variance

$$\hat{V}(\bar{A}_n) = \sum_{h=1}^L W_h^2 n_h^{-1} (n_h - 1)^{-1} \sum_{k=1}^{n_h} (A_{hk} - \bar{A}_h)^2. \tag{4.5}$$

Table 5
Estimated Cotton Acreages from 500 Simulated
June Enumerative Survey Samples

	\bar{A}_n	$\hat{V}(\bar{A}_n)$
Average	9.93	0.64
Range	8.13 – 12.21	
Variance	0.66	0.016

The average cotton acres per segment in the simulated population was 9.94, while the average of the 500 sample estimates was 9.93. The actual variance of the stratified estimator \bar{A}_n was 0.63, while the average estimated variance for the 500 simulated samples was 0.64. Because the variation in estimated cotton acreage is small, π_{hk}^* provides a stable estimate of the unconditional probability that segment k in stratum h is selected to receive at least one secondary unit.

In addition to the estimators discussed previously, random group estimators of the variance were constructed. Two sets of random groups were formed for each objective yield survey sample. One set contained five groups ($\gamma = 5$) and one set contained ten groups ($\gamma = 10$). Random groups were created by dividing the primary sampling units, the segments, into subsets within each land use substratum. The first group in each set of groups was obtained by drawing a simple random sample without replacement of size $K_{h(\gamma)} = n_h/\gamma$ from the sample of segments selected from each stratum ($h=1, \dots, 28$) of the parent June Enumerative Survey sample. The second random group was obtained in the same fashion by selecting $K_{h(\gamma)}$ segments from the remaining $n_h - K_{h(\gamma)}$ segments in each stratum. The remaining random groups were formed in a like manner. One land use substratum, stratum number 3107, had a sample size of $n_h = 5$ segments. Acreage and yield values of the observed five segments were repeated to form the ten observations required to create ten groups when $\gamma = 10$.

Let D_α be the number of secondary units with positive acres which were selected during the objective yield survey in random group α where $\alpha = 1, \dots, \gamma$. Let $\bar{y}_{(\alpha)}$ denote the yield estimator obtained from the α -th random group:

$$\bar{y}_{(\alpha)} = D_\alpha^{-1} \sum_{t=1}^{D_\alpha} Y_{t(\alpha)}, \tag{4.6}$$

where $\bar{y}_{(\alpha)}$ is the analogue of equation (3.3) for the α -th group. The random group estimator of the variance of \bar{y} is then given by

$$\hat{V}_{g\gamma}(\bar{y}) = \gamma(\gamma - 1)^{-1} \sum_{\alpha=1}^{\gamma} (\bar{y}_{(\alpha)} - \bar{y})^2. \tag{4.7}$$

This estimator is slightly biased for the ten group estimator because one stratum contained only five observations, and these observations were repeated in the groups.

Similarly, let $\hat{Y}_{(\alpha)}$ denote the total production estimator obtained from the α -th random group:

$$\hat{Y}_{(\alpha)} = N \bar{M}_{n(\alpha)} \bar{y}_{(\alpha)}, \tag{4.8}$$

where

$$\bar{M}_{n(\alpha)} = \sum_{h=1}^L w_h K_{h(\alpha)}^{-1} \sum_{k=1}^{K_{h(\alpha)}} M_{hk(\alpha)},$$

$M_{hk(\alpha)}$ is the number of acres of cotton in segment k of stratum h for random group α and $K_{h(\alpha)}$ is the number of segments in stratum h for the α -th group. The random group estimator of the variance of \hat{Y} is then given by

$$\hat{V}_{g\gamma}(\hat{Y}) = \gamma(\gamma - 1)^{-1} \sum_{\alpha=1}^{\gamma} (\hat{Y}_{(\alpha)} - \hat{Y})^2. \tag{4.9}$$

Tables 6 and 7 summarize the results of the Monte Carlo study for yield and production estimators. Average values of the estimates and their variance estimates across the 1,000 simulated objective yield survey samples are shown in the tables. Simulation of two objective yield survey samples for each June Enumerative Survey sample made the estimation of between – and within – June Enumerative Survey variance components possible.

The estimator (3.1) currently used, \bar{y} , and the combined ratio estimator (3.15), \bar{y}_r , which is based on the π_{hk}^* calculated from June Enumerative survey results, provide estimates with similar accuracy (see Table 6). The equal efficiency is partly due to the accuracy with which the unconditional selection probabilities are estimated in each sample.

As was shown in Section 3.2, the conditional variance $\hat{V}_2(\bar{y})$ is an underestimate of $V(\bar{y})$. For this simulated population, $\hat{V}_2(\bar{y})$ underestimated the observed variance of \bar{y} by 38%. The observed variance of \bar{y} was 11.57 as compared to an average of 7.21 for $\hat{V}_2(\bar{y})$. This underestimation of the variance was consistent across samples. The estimated variance of $\hat{V}_2(\bar{y})$ was 0.99, with $\hat{V}_2(\bar{y})$ ranging from a low of 3.85 to a high of 11.24 in the 1,000 observations. Thus, the maximum observed estimate of the conditional variance was less than the true variance.

Table 6
Monte Carlo Properties of Yield per Acre Estimates
and Estimated Variances¹

	Estimator					
	\bar{y}	$\hat{V}_2(\bar{y})$	$\hat{V}_{g5}(\bar{y})$	$\hat{V}_{g10}(\bar{y})$	\bar{y}_r	$\hat{V}(\bar{y}_r)$
Average	79.74	7.21	12.62	12.39	79.76	12.39
Total Variance	11.57	0.99	74.58	36.86	11.56	12.51
Between JES	7.60	0.48	6.10	4.56	7.64	7.61
Within JES	3.97	0.51	68.48	32.30	3.92	4.90

¹ Two objective yield survey samples were simulated from each of 500 simulated June Enumerative Survey samples.

Table 7
Monte Carlo Properties of Production Estimates
and Estimated Variances¹

	Estimator ²					
	\hat{Y}	$\hat{V}_*(\hat{Y})$	$\hat{V}_{g^5}(\hat{Y})$	$\hat{V}_{g^{10}}(\hat{Y})$	\hat{Y}_r	$\hat{V}(\hat{Y}_r)$
Average	73.04	40.85	48.99	48.53	73.07	48.73
Total Variance	49.69	82.52	1245.10	608.80	49.58	222.96
Between JES	46.35	78.17	50.82	208.48	46.30	199.58
Within JES	3.34	4.35	1194.28	400.32	3.28	23.38

¹ Two objective yield survey samples were simulated from each of 500 simulated June Enumerative Survey samples. There were $N = 92,240$ segments in the simulated population.
² The estimator \hat{Y} is in millions of units and variances are in the corresponding units.

Assuming probability-proportional-to-size sampling with replacement of segments at the second phase, $\hat{V}_2(\bar{y})$ was shown in Section 3.2 to be unbiased for the variance of \bar{y} conditional on the sample of segments selected at the first stage of sampling. The estimate of the expected value of the conditional variance of \bar{y} , $V_2(\bar{y})$, from the Monte Carlo study is 3.97. This large discrepancy (3.97 versus 7.21) can be attributed to the fact that the estimator $\hat{V}_2(\bar{y})$ ignores the effects of stratification in the population (see Tables 2 and 3) and to the fact that $\hat{V}_2(\bar{y})$ was derived under the assumption that segments are selected with replacement at the second stage of sampling.

The estimator (3.9), $\hat{V}_*(\hat{Y})$, underestimates the unconditional variance of \hat{Y} . While the observed variance of \hat{Y} from the Monte Carlo simulations is 49.69 (million)², the average of the $\hat{V}_*(\hat{Y})$ is only 40.85 (million)². This 18% underestimate of the true variance occurs for a number of reasons. As was shown previously, there is a negative bias in $\hat{V}_2(\bar{y})$ as an estimator of $\hat{V}(\bar{y})$; another important factor contributing to the bias is the failure of $\hat{V}_*(\hat{Y})$ to take into account the covariance between \bar{M}_n and \bar{y} . In this example, the bias caused by omitting the covariance term partially balances the bias associated with $\hat{V}(\bar{y})$.

Using expression (3.16), $\hat{V}(\bar{y}_r)$, as an estimator of the variance of \bar{y}_r and expression (3.21), $\hat{V}(\hat{Y}_r)$, as an estimator of the variance of \hat{Y}_r provided results which are much more satisfactory than those of the estimators currently used. The Monte Carlo average of the estimates $\hat{V}(\bar{y}_r)$ was 12.51, which overestimates the observed variance of \bar{y}_r (11.57) by about 7%. About one-third of the overestimate (2-4%) can be attributed to the use of sampling without replacement at the first two stages of sampling. The remaining difference of about 4% is small relative to the standard error of the estimated difference. The variance of the difference was estimated by estimating the variance of the mean of z_{ij} , where

$$z_{ij} = (\bar{y}_{n,r(tj)} - 79.76)^2 - \hat{V}(\bar{y}_{n,r(tj)}), \tag{4.10}$$

for the j -th yield sample ($j = 1, 2$) within June Enumerative Survey sample t ($t = 1, \dots, 500$). The estimated standard error of the difference was 0.58. Thus, the average value of $\hat{V}(\bar{y}_r)$ is within 1.5 standard errors of the estimated variance of \bar{y}_r . The average estimated variance of \hat{Y}_r is within 2 percent of the variance observed in the Monte Carlo simulations.

Random group estimators of the variance of \bar{y} displayed little bias. The Monte Carlo averages of estimators $\hat{V}_{g5}(\bar{y})$ and $\hat{V}_{g10}(\bar{y})$ were 9% and 7%, respectively, larger than the corresponding Monte Carlo variances. These differences are not significantly different from zero and are comparable to those obtained for the estimator $\hat{V}(\bar{y}_r)$. The variance estimator $\hat{V}(\bar{y}_r)$, however, is a much more stable variance estimator. The coefficient of variation for the estimator $\hat{V}(\bar{y}_r)$ is about 30%; it is 75% for $\hat{V}_{g5}(\bar{y})$. As expected (Wolter 1985), an increase in the number of random groups resulted in a decrease in the coefficient of variation of the random group variance estimator. The coefficient of variation for $\hat{V}_{g10}(\bar{y})$ was 50%. Differences among random groupings and yield samples within June Enumerative Surveys accounted for most of the variance in the random groups variance estimators.

4. CONCLUSIONS

Analyses show that the estimators of statewide average yield and total production currently used by the National Agricultural Statistics Service are satisfactory. However, the simple variance estimators $\hat{V}_2(\bar{y})$ and $\hat{V}_*(\hat{Y})$ were shown to have a negative bias, where the extent of the underestimation is a function of the within-segment variance and of the within-segment sampling rates. The estimator $\hat{V}_2(\bar{y})$ underestimated the true variance of \bar{y} by nearly 40%, and $\hat{V}_*(\hat{Y})$ underestimated the true variance of \hat{Y} by 18% for the simulated California cotton population.

The alternative estimators, \bar{y}_r and \hat{Y}_r , were developed by viewing the yield sampling scheme as a two-phase process in which segments found to contain crop acreage during phase one (the June Enumerative Survey) are subsampled during phase two to estimate yield. The unconditional probability of selecting a segment to receive a secondary unit within a stratum, π_{hk}^* , is estimated by assuming that this probability is proportional to the conditional probability of selecting segments at the second phase of sampling. With this assumption, the unequal probability combined ratio estimator of the mean yield, \bar{y}_r , and the estimator of its variance, $\hat{V}(\bar{y}_r)$, were developed. The estimator of the total \hat{Y}_r is a two-phase product estimator of the mean production per segment, where the estimator of the mean of the auxiliary variable (crop acreage) comes from the June Enumerative Survey (phase one of sampling). The variance estimator $\hat{V}(\hat{Y}_r)$ is a stratified double sampling (two-phase) estimator of the variance of \hat{Y}_r .

As shown by the Monte Carlo study, \bar{y}_r and \hat{Y}_r give estimates that are comparable to their currently used counterparts, \bar{y} and \hat{Y} . Both $\hat{V}(\bar{y}_r)$ and $\hat{V}(\hat{Y}_r)$ are accurate variance estimators in samples of the size typically used by NASS. These results are due, in part, to the precision with which average crop acreages are estimated by the June Enumerative Survey. Precise acreage estimates produce estimates of selection probabilities that are close to the unconditional probabilities of selection. In addition, the ratio form of the estimator reduces the effect of replacing true unconditional probabilities with estimators.

Random group variance estimators are also essentially unbiased estimators of the variance of estimated yield and production. However, random group estimators are much less stable than $\hat{V}(\bar{y}_r)$ and $\hat{V}(\hat{Y}_r)$. Therefore, estimators $\hat{V}(\bar{y}_r)$ and $\hat{V}(\hat{Y}_r)$ are recommended over random group estimators.

The June Enumerative Survey forms phase one of the objective yield survey. Sampling procedures for the June Enumerative Survey are straightforward and, as was shown by the Monte Carlo study, provide accurate acreage estimates. Hence, no change in the overall design for phase one of the objective yield survey is recommended.

A number of modifications for phase two of the objective yield surveys should be investigated. The current procedure for estimating yield is a two phase procedure in which a combined ratio estimator is used. In states where the sample is relatively large, independent sampling at phase two within individual strata or for groups of strata, as well as the use of a separate ratio estimator should be considered.

Systematic sampling at phase two should be replaced if unbiased estimators of the variance are desired. Segments for yield sampling at phase two are now selected by computer at a national level so it should be relatively easy to change to a selection procedure with known joint selection probabilities. Estimators similar to those recommended for the current design would still be suitable if the same selection probabilities were retained. The scheme described by Fuller (1970) is one procedure that can be computerized, for which joint selection probabilities can be calculated, and which maintains specified selection probabilities and a degree of control similar to that of systematic sampling.

ACKNOWLEDGEMENTS

This research was supported in part by cooperative research agreement 58-319T-1-0054X with the National Agricultural Statistics Service, U.S. Department of Agriculture. We thank the referees for useful comments.

REFERENCES

- FECISO, R. (1978). Cluster analysis as an aid in creating paper strata. Statistical Reporting Service, U.S. Department of Agriculture.
- FECISO, R., and JOHNSON, V. (1981). The new California area frame: A statistical study. Statistical Reporting Service, U.S. Department of Agriculture.
- FULLER, W.A. (1970). Sampling with random stratum boundaries. *Journal of the Royal Statistical Society*, Ser. B, 32, 209-226.
- HOUSEMAN, E.E. (1975). Area frame sampling in agriculture. Statistical Reporting Service, U.S. Department of Agriculture.
- PRATT, W.L. (1984). The use of interpenetrating sampling in area frames. Statistical Reporting Service, U.S. Department of Agriculture.
- WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

Current Issues on Seasonal Adjustment

ESTELA BEE DAGUM¹

ABSTRACT

This paper discusses three problems that have been a major preoccupation among researchers and practitioners of seasonal adjustment in statistical bureaus for the last ten years. These problems are: (1) the use of concurrent seasonal factors versus seasonal factor forecasts for current seasonal adjustment; (2) finding an optimal pattern of revisions for series seasonally adjusted with concurrent factors; and (3) smoothing highly irregular seasonally adjusted data.

KEY WORDS: Concurrent vs forward seasonal factors; Revisions; Trend-cycle filters; Smoothing.

1. INTRODUCTION

During the last decade, within the domain of seasonal adjustment, statistical bureaus have focused their attention on three important issues: (1) the seasonal adjustment of a current value; (2) the revisions of concurrent seasonally adjusted data; and (3) the smoothing of highly irregular seasonally adjusted series.

The main purpose of this article is to discuss each of the above problems with respect to the X-11-ARIMA seasonal adjustment program developed by Dagum (1980) and which is applied by Statistics Canada and other statistical bureaus of the world.

The four modes in which the X-11-ARIMA computer package can be used to produce a current seasonally adjusted value are discussed in Section 2. In Section 3, the focus is on analysis of the revisions of concurrent seasonally adjusted data based on the linear filters of X-11-ARIMA. Section 4 deals with the nature and characteristics of the smoothing (trend-cycle) filters available in X-11-ARIMA.

2. SEASONAL ADJUSTMENT OF CURRENT VALUES

The seasonal adjustment of a current value can be done using either a "concurrent" seasonal estimate or a seasonal "forecast".

A "concurrent" seasonal estimate (factor or effect depending on whether a multiplicative or additive model is assumed) is obtained by seasonally adjusting, each time a new observation is available, all the data available up to and including that observation. On the other hand, a seasonal "forecast" is obtained from a series that ended in the previous year. A common practice is to generate these seasonal forecasts, say for year $t + 1$, from data that ended in December of the previous year t .

There are four modes in which the X-11-ARIMA computer program can be applied to produce a current (last observation) seasonally adjusted value. These four modes are: (i) using ARIMA extrapolations and concurrent seasonal factors; (ii) using ARIMA extrapolations and seasonal factor forecasts; (iii) using concurrent seasonal factors without the use of ARIMA extrapolations; and (iv) using seasonal factor forecasts without the use of ARIMA extrapolations.

¹ Estela Bee Dagum, Time Series Research and Analysis Division, Methodology Branch, Statistics Canada, 13th Floor, R.H. Coats Building, Ottawa, Ontario, Canada K1A 0T6.

While statistical bureaus use the four modes to obtain current seasonally adjusted values, not all of them do so with the same frequency. Thus, for example, the dominant mode in Statistics Canada is (i) followed by mode (iii) whereas in the U.S. Bureau of Labor, the dominant mode is (ii) followed by mode (iv). The current seasonally adjusted value produced by each type of seasonal adjustment varies and is subject to different degrees of error.

Under the assumption of an additive decomposition model, the seasonal adjustment of a current value X_t can be obtained by

$$\hat{X}_t^{(\ell)} = X_t - \hat{S}_t^{(\ell)}, \quad (1)$$

where $\hat{S}_t^{(\ell)}$ denotes a forward seasonal estimate; or by

$$\hat{X}_t^{(0)} = X_t - \hat{S}_t^{(0)}, \quad (2)$$

where $\hat{S}_t^{(0)}$ denotes a concurrent seasonal estimate.

The current seasonally adjusted value will become "final" in the sense that it will no longer be revised after m more observations are added. Thus,

$$\hat{X}_t^{(m)} = X_t - \hat{S}_t^{(m)}, \quad (3)$$

where $\hat{S}_t^{(m)}$ denotes a final seasonal estimate.

Therefore, the total revision of a concurrent and of a forward seasonal estimate can be written as

$$r_t^{(0,m)} = \hat{S}_t^{(0)} - \hat{S}_t^{(m)}, \quad m > 0; \quad (4)$$

$$r_t^{(\ell,m)} = \hat{S}_t^{(\ell)} - \hat{S}_t^{(m)}, \quad m > 0 > \ell. \quad (5)$$

Under the assumption of an additive decomposition and no replacement of extreme values, $\hat{S}_t^{(m)}$, the final seasonal estimate from a series $X_{t-m}, \dots, X_t, \dots, X_{t+m}$ can be expressed by

$$\hat{S}_t^{(m)} = \sum_{j=-m}^m h_{m,j} X_{t-j} = h^{(m)}(B) X_t, \quad (6)$$

where $h_{m,j} = h_{m,-j}$ are the symmetric moving average weights to be applied to the series. $h^{(m)}(B)$ denotes the corresponding linear filter using the backshift operator B , such that $B^n = X_{t-n}$. Young (1968) showed that the length of this symmetric filter $h^{(m)}(B)$, for monthly series, is 145 but that it can be well approximated by 85 weights because the values of the weights attached to distant observations are very small and, thus, $m = 42$.

Following equation (6) we can express a concurrent seasonal estimate $\hat{S}_t^{(0)}$ and a forward seasonal estimate $\hat{S}_t^{(\ell)}$ by:

$$\hat{S}_t^{(0)} = \sum_{j=-2m}^0 h_{0,j} X_{t-j} = h^{(0)}(B) X_t, \quad m = 42, \quad (7)$$

where $h^{(0)}(B)$ denotes the asymmetric *concurrent* seasonal filter; and

$$\hat{S}_t^{(\ell)} = \sum_{j=-2m}^{\ell} h_{\ell,j} X_{t-j} = h^{(\ell)}(B) X_t, \quad m = 42, \tag{8}$$

where $h^{(\ell)}(B)$ denotes the asymmetric forecasting seasonal filter and $\ell = 1, 2, \dots, 12$ for a monthly series.

The revision of a concurrent seasonal estimate depends on the distance between the concurrent and the final filter, that is, $d[h^{(0)}(B), h^{(m)}(B)]$, and on the innovations of the new observations $X_{t+1}, X_{t+2}, \dots, X_{t+m}$.

Similarly, the revision of a forward seasonal estimate depends on $d[h^{(\ell)}(B), h^{(m)}(B)]$ and on the new innovations introduced by $X_{t-\ell}, \dots, X_t, X_{t+1}, \dots, X_{t+m}$.

Theoretical studies by Dagum (1982a and 1982b) have shown that

$$d[h^{(0)}(B), h^{(m)}(B)] < d[h^{(\ell)}(B), h^{(m)}(B)] \text{ for } \ell = 1, 2, \dots, 12. \tag{9}$$

The distance between the two filters is defined as the mean squared difference between the frequency response function of the filters over all the seasonal frequencies; a similar definition is given in the next section (equation (17)) using the root mean squared difference.

Relation (9) is true whether ARIMA extrapolations are used or not. Furthermore, the two studies also showed that

$$\begin{aligned} & d[h^{(0)}(B), h^{(m)}(B)] \text{ using ARIMA extrapolations} \\ & < d[h^{(0)}(B), h^{(m)}(B)] \text{ without ARIMA extrapolations,} \end{aligned} \tag{10}$$

and similarly

$$\begin{aligned} & d[h^{(\ell)}(B), h^{(m)}(B)] \text{ using ARIMA extrapolations} \\ & < d[h^{(\ell)}(B), h^{(m)}(B)] \text{ without ARIMA extrapolations,} \end{aligned} \tag{11}$$

$$\text{for } \ell = 1, 2, \dots, 12.$$

Studies by Dagum (1978), Bayer and Wilcox (1981), Kenney and Durbin (1982), McKenzie (1984), Dagum and Morry (1984), Pierce (1980) and Pierce and McKenzie (1985) have shown that

$$r^{(0,m)} < r^{(\ell,m)} \tag{12}$$

except in a few cases where

$$r^{(0,m)} > r^{(\ell,m)}. \tag{13}$$

The relationship (13) can be observed when the current observations of the latest year are strongly revised since X_t gets the largest weight in the estimations of $\hat{S}_t^{(0)}$.

From the viewpoint of the total revisions of the seasonal estimates, the results of the above empirical studies permit the ranking of the four modes as follows: mode (i) (ARIMA extrapolations with concurrent seasonal estimates) gives the smallest total revision; mode (iii) (no ARIMA extrapolations with concurrent seasonal estimates) ranks second; mode (ii) (ARIMA extrapolations with forward seasonal estimates) ranks third and mode (iv) (ARIMA extrapolations with forward seasonal estimates) ranks fourth.

3. REVISIONS OF CONCURRENT SEASONALLY ADJUSTED DATA

Statistics Canada's practice of using concurrent seasonal adjustment was first established in 1975 for the Labour Force Survey series. Gradually other foreign statistical agencies followed it. The use of concurrent seasonal factors for current seasonal adjustment poses the problem of how often should the series be revised. Kenny and Durbin (1982) recommended that revisions should be made after one month and thereafter each calendar year. Dagum (1982c) supported these conclusions and furthermore, recommended an additional revision at six months if the seasonal adjustment method is the X-11-ARIMA without the ARIMA extrapolation option.

For any two points in time $t + k$, $t + \ell$ ($k < \ell$), the revisions of the seasonal estimates and consequently of the seasonally adjusted value is given by

$$r_t^{(\ell, k)} = \hat{X}_t^{(\ell)} - \hat{X}_t^{(k)}, \quad k < \ell. \quad (14)$$

This revision reflects: (1) the innovations introduced by the new observations X_{t+k+1} , X_{t+k+2} , ..., $X_{t+k+\ell}$; and (2) the differences between the two asymmetric seasonal adjustment filters $Y^{(\ell)}(B)$ and $Y^{(k)}(B)$. If one fixes $k = 0$ and lets ℓ vary from 1 to m , then relation (14) gives a sequence of revisions of the concurrent seasonally adjusted values for different time spans or lags. The *total revision* of the concurrent estimate is given for $\ell = m$. If one fixes $\ell = k + 1$ and lets k take values from 0 to $m - 1$, then relation (14) gives the sequence of *single period revisions* of each estimated seasonally adjusted value and in particular, if one starts at $k = 0$ one obtains the $m - 1$ successive single period revisions of each estimated seasonally adjusted value before it becomes final. If one fixes $\ell = k + 12$ and lets k take values from 0 to $m - 12$, then equation (14) gives the sequence of annual revisions.

The revisions in which we are interested here are those introduced by filter discrepancies, and these can be studied by looking at the frequency response functions of the corresponding filters. Similarly to equation (6), we can approximate the seasonally adjusted value for recent years from the X-11-ARIMA program (with or without ARIMA extrapolations) by

$$\hat{X}_t^{(n)} = \sum_{j=n}^m Y_{n,j} X_{t-j} = Y^{(n)}(B) X_t. \quad (15)$$

Equation (15) represents a linear system where $\hat{X}_t^{(n)}(n)$ is the convolution of the input X_t and a sequence of weights $Y_{n,j}$ called the *impulse response function* of the filter. The properties of this function can be studied using its Fourier transform which is called the *frequency response function*, defined by

$$\Gamma^{(n)}(\omega) = \sum_{j=-n}^m Y_{n,j} e^{-2\pi\omega j}, \quad -1/2 \leq \omega \leq 1/2, \quad (16)$$

where ω is the frequency in cycles per unit time. $\Gamma^{(n)}(\omega)$ fully describes the effects of the linear filter on the given input. Monthly and annual revisions of the concurrent filter of X-11-ARIMA with and without the ARIMA extrapolations have been calculated by Dagum (1987) based on the mathematical distance between the various frequency response functions of the filters. The pattern is characterized by a rapid decrease in the size of the monthly revisions of the concurrent filter for $\ell = 1, 2$, and 3; and a slow decrease thereafter until $\ell = 11$; then a large increase occurs at $\ell = 12$ followed by a decrease at $\ell = 13$ and then another large increase at $\ell = 24$ followed by a decrease at $\ell = 25$. Dagum (1987) showed that this pattern of monthly revisions is the same whether ARIMA extrapolations are used or not.

The significant decreases for the first three consecutive revisions are due to the improvement of the Henderson (trend- cycle) filter weights. The reversal of direction in the size of the filter revisions at $\ell = 12$ and $\ell = 13$, is due to the improvements of the seasonal filter that becomes less asymmetrical from year to year until three full years are added to the series. The two largest revisions occur at $\ell = 1$ and $\ell = 12$. *Given the non-monotonicity of single monthly revisions, it is not advisable to revise the concurrent estimate any time a new observation is added to the series.*

A revision scheme often used by statistical bureaus for their concurrent seasonally adjusted series consists of keeping constant the concurrent estimate from the time it appears until the end of the year and then revising annually the current and earliest years. Therefore, first year revisions due to filter discrepancies are given by $R^{(0,0)}$, $R^{(1,0)}$, ..., $R^{(11,0)}$; second year revisions by $R^{(12,0)}$, $R^{(13,1)}$, ..., $R^{(23,11)}$ third-year revisions by $R^{(24,12)}$ $R^{(25,13)}$ and so on where $R^{(\ell,k)}$ is defined by

$$R^{(\ell,k)} = [2\int_0^{1/2} \|\Gamma^{(\ell)}(\omega) - \Gamma^{(k)}(\omega)\|^2 d\omega]^{1/2}, \tag{17}$$
$$\ell = 1, 2, \dots, n, k = 0, 1, 2, \dots, n - 12,$$

and $n = 42$ for the X-11-ARIMA seasonal adjustment filters.

Table 1 shows the first-, second- and third-year revisions of the concurrent seasonal adjustment filter for X-11-ARIMA without extrapolation and with extrapolations from one ARIMA model and two sets of parameter values (other cases are shown in Dagum 1987). The ARIMA model chosen is the classical (0,1,1) (0,1,1)₁₂ model that is $(1 - B)(1 - B^{12})X_t = (1 - \theta B)(1 - \Theta B^{12})a_t$ where X_t denotes the original series, B is the backshift operator such that $B^n X_t = X_{t-n}$, a_t is a purely random process that represents the innovations and θ and Θ are the non-seasonal and seasonal parameters, respectively.

Since the largest single period revisions occur at $\ell = 1$ and $\ell = 12$ as mentioned above, a better revision scheme would be to incorporate monthly and annual revisions. It is expected that (1) adjusting concurrently each month, say from January to November and revising only once when the next month is available, and (2) adjusting concurrently December when it first appears and then revising the first year and earlier years when January is added, should improve the reliability of the filter applied during the current year while maintaining simultaneously the filter's homogeneity for month-to-month comparisons.

The first-year revisions of the first-month revised filter would then be $R^{(1,1)}$, $R^{(2,1)}$, ..., $R^{(11,1)}$. Table 2 shows these revisions and although the pattern is very similar to that of the concurrent filter, *the size of the revisions are much smaller if no extrapolations are used*. On the other hand, *the improvement is less important if ARIMA extrapolations are used*. Similarly, no major differences were observed for the second- and third-year revisions.

3.1 Estimation of Trading Day Variations and ARIMA Models with Concurrent Seasonal Adjustment

Besides the type of revisions scheme to be applied, there are two other problems posed by concurrent seasonal adjustment associated with trading day variations and ARIMA modelling.

Table 1
First-, Second- and Third-Year Revisions of the Concurrent
Seasonal Adjustment Filter of X-11-ARIMA

Revisions $R^{(\ell,k)}$	Without ARIMA Extrapolations	With ARIMA Extrapolations from a (0,1,1) (0,1,1) ₁₂ Model			
		$\theta = .40$	$\Theta = .80$	$\theta = .80$	$\Theta = .80$
$R^{(1,0)}$.12	.12		.06	
$R^{(2,0)}$.13	.13		.08	
$R^{(3,0)}$.13	.13		.08	
$R^{(4,0)}$.13	.13		.09	
$R^{(5,0)}$.15	.13		.09	
$R^{(6,0)}$.17	.13		.09	
$R^{(7,0)}$.16	.13		.09	
$R^{(8,0)}$.16	.13		.09	
$R^{(9,0)}$.16	.13		.09	
$R^{(10,0)}$.16	.14		.09	
$R^{(11,0)}$.16	.14		.09	
$R^{(12,0)}$.29	.28		.26	
$R^{(13,1)}$.27	.27		.26	
$R^{(14,2)}$.27	.27		.26	
.	.	.		.	
.	.	.		.	
$R^{(23,11)}$.27	.26		.26	
$R^{(24,12)}$.20	.16		.16	
$R^{(24,13)}$.18	.17		.16	
$R^{(36,24)}$.16	.17		.16	
.	.	.		.	
.	.	.		.	
.	.	.		.	

Table 2
First-Year Revisions of the First-Month Revised
Seasonal Adjustment Filter

Revisions $R^{(\ell,1)}$	Without ARIMA Extrapolations	With ARIMA Extrapolations from a (0,1,1) (0,1,1) ₁₂ Model			
		$\theta = .40$	$\Theta = .80$	$\theta = .80$	$\Theta = .80$
$R^{(2,1)}$.07	.10		.06	
$R^{(3,1)}$.07	.10		.06	
$R^{(4,1)}$.07	.10		.07	
$R^{(5,1)}$.08	.10		.08	
$R^{(6,1)}$.10	.11		.08	
$R^{(7,1)}$.11	.11		.08	
$R^{(8,1)}$.11	.11		.08	
$R^{(9,1)}$.11	.11		.08	
$R^{(10,1)}$.12	.11		.08	
$R^{(11,1)}$.12	.12		.08	

For series which are flows in the sense that they result from the accumulation of daily values over the calendar months, there is a systematic effect caused by trading day variations. Trading day variations arise mainly because the activity varies with the days of the week. Other sources are associated with accounting and reporting practices. For example, stores that do their bookkeeping activities on Friday tend to report higher sales in months with five Fridays than in months with four Fridays. The trading day effects are estimated in the X-11-ARIMA program using ordinary least squares on a simple deterministic regression model. Consequently, the weights estimated for each day change any time a new observation is added to the series. Since regression techniques are very sensitive to outliers, these changes can be sometimes unnecessarily large.

When the series are seasonally adjusted concurrently, the trading day estimates change all the time. In order to avoid unnecessary revisions, Statistics Canada's practice is to use the weights calculated by the program at the end of the previous calendar year or the weights provided by the users, as priors for the current year. The weights are then revised on an annual basis.

The effect of trading day variations must be removed from the series before ARIMA modelling, for these type of models cannot adequately handle trading day variations. In other words, if the X-11-ARIMA program is used with ARIMA extrapolations on series with trading day variations, these variations should be estimated *a priori* and if significant, they should be removed from the original series before the ARIMA modelling.

Another problem associated with concurrent seasonal adjustment refers to how often the ARIMA models should be identified. The current practice at Statistics Canada is to use the automatic ARIMA model selection option once a year and if the model is accepted, then it is kept constant for a whole year, letting only the parameters change when more observations are added. In order to keep the model constant, the user's supplied model option should be applied. Maintaining the ARIMA model constant avoids unnecessary revisions that may result from changing of models back and forth simply because of the presence of outliers.

4. SMOOTHING OF VOLATILE SEASONALLY ADJUSTED SERIES

One of the main purposes of the seasonal adjustment of economic time series is to provide information on current economic conditions, particularly to determine the stage of the cycle at which the economy stands. Since seasonal adjustment means removing only seasonal variations, thus leaving trend-cycle variations together with irregular fluctuations, it is often difficult to detect the short-term trend or cyclical turning points for series strongly affected with irregulars. In such cases, it may be preferable to smooth the seasonally adjusted series using trend-cycle estimators which suppress as much as possible the irregulars without affecting the cyclical component.

The use of trend-cycle values has been discussed by several writers and recently by Moore *et al* (1981), Kenny and Durbin (1982), Maravall (1986) and Dagum and Laniel (1987). Although not yet practised widely, some statistical agencies such as Statistics Canada and the Australian Bureau of Statistics smooth some of their seasonally adjusted series, particularly those series that are strongly affected by irregulars.

The combined linear filters applied to the original series to generate a central (symmetric) estimate of the trend-cycle component have been calculated by Young (1968) for Census Method II-X-11 variant. This filter is similar to that of X-11-ARIMA with and without ARIMA extrapolations. Dagum and Laniel (1987) extended Young's (1968) results to include the estimation of the asymmetric trend-cycle filters of X-11-ARIMA with and without the ARIMA extrapolations.

Figure 1 shows the gain functions of the central (symmetric) seasonal adjustment filters and smoothed seasonally adjusted data (trend-cycle) filters. It is apparent that the trend-cycle filters suppress all the noise present in the series, where the noise is defined as the power present in all frequencies $\omega \leq .166$. This frequency corresponds to the first harmonic of the fundamental seasonal frequency of a monthly series. This pattern results from the convolution of the seasonal adjustment filters with the 13-term Henderson trend-cycle filter.

Figure 2a shows the gain functions of the concurrent and first-month revised trend-cycle filters of X-11-ARIMA *without* ARIMA extrapolations. Figure 2b shows their corresponding phase-shift functions expressed in months instead of radians. We can observe that the gain for all $\omega \leq .166$ is much larger for these two asymmetric filters as compared with the central filter. Furthermore, there are large amplifications for frequencies near the fundamental seasonal. All this means that the concurrent and first revised smoothed seasonally adjusted values will have more noise than the final estimates. On the other hand, it is apparent that the phase shifts are very small, less than one month for the most important cyclical frequencies $0 < \omega < .055$ (i.e., cycles of periodicities equal to and longer than 18 months).

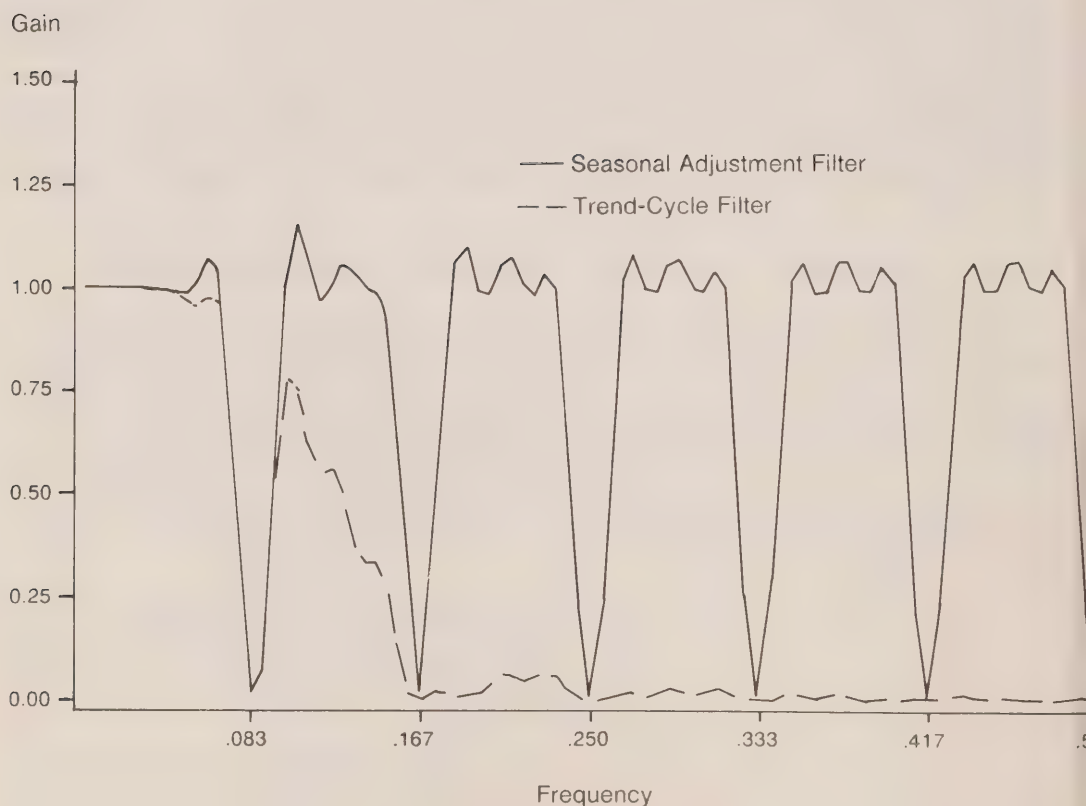


Figure 1. Gain Functions of the Central (Symmetric) Trend-Cycle and Seasonal Adjustment Filters of X-11-ARIMA.

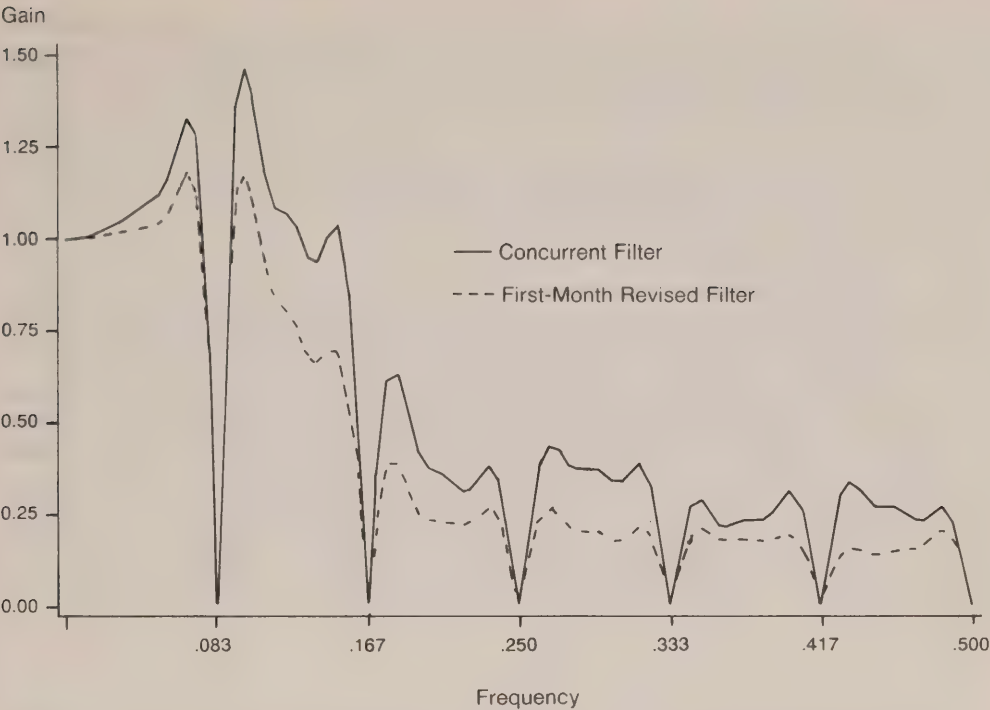


Figure 2a. Gain Functions of the Concurrent and First-Month Revised Filters of X-11-ARIMA without ARIMA Extrapolations.

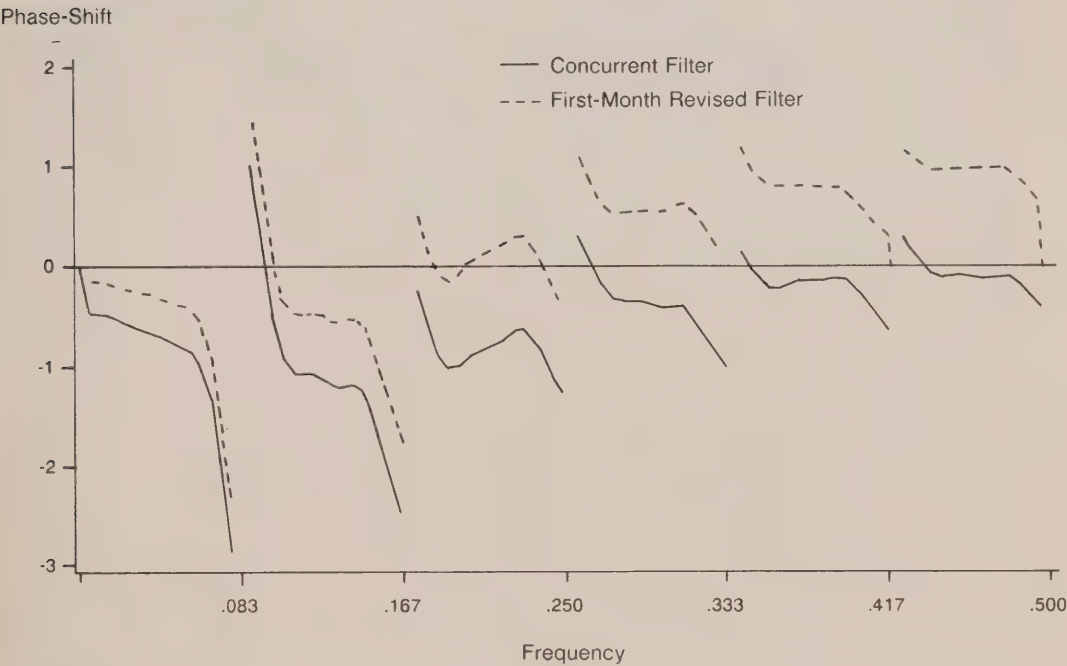


Figure 2b. Phase-Shift Functions of the Concurrent and First-Month Revised Filters of X-11-ARIMA without ARIMA Extrapolations.

Figures 3a and 3b show the gain and phase-shift functions of the concurrent and first-month revised trend-cycle filters of X-11-ARIMA with ARIMA extrapolations. The extrapolations are obtained from an IMA model $(0,1,1)(0,1,1)_{12}$ with $\theta = .40$ and $\Theta = .60$. The gain functions are closer to the symmetric (central) filter than those of X-11-ARIMA without the ARIMA extrapolations. There are no amplifications around the fundamental seasonal frequency and a similar attenuation of power at higher frequencies. On the other hand, there is more phase-shift (being near to one month) for low frequencies and less phase-shift for all high frequencies.

Dagum and Laniel (1987) studied the time path of the revisions of the trend-cycle filters and compared them with those of the seasonal adjustment filters. Their results, as summarized in Table 3, show that the total revisions of the trend-cycle asymmetric filters converge to zero much faster than those of the corresponding seasonal adjustment filters. In fact, the total revision of the trend-cycle filter three months after the concurrent filter is only .1, whereas a close value is achieved for the seasonal adjustment filter only after 24 months have been added to the series. Except for the total revisions of the concurrent filter which is larger for the trend-cycle filters compared with the corresponding seasonal adjustment filter, in all the other cases the total revisions are smaller for the trend-cycle filters. Furthermore, the trend-cycle filter revisions converge much faster to zero as compared with those of the seasonal adjustment filters.

Table 3
Time Path of the Total Revisions of the Trend-Cycle and the Seasonal Adjustment
Asymmetric Filters of X-11-ARIMA

Revisions $R^{(\ell,k)*}$	Without Extrapolations		With Extrapolations from a $(0,1,1)(0,1,1)_{12}$ Model $\theta = .40 \quad \Theta = .60$	
	Trend-Cycle Filter	Seasonal Adjustment Filter	Trend-Cycle Filter	Seasonal Adjustment Filter
$R^{(48,0)}$.45	.36	.41	.32
$R^{(48,1)}$.27	.33	.26	.32
$R^{(48,2)}$.15	.32	.15	.32
$R^{(48,3)}$.11	.32	.11	.31
$R^{(48,4)}$.12	.32	.11	.31
.
.
.
$R^{(48,12)}$.10	.23	.09	.20
$R^{(48,24)}$.07	.13	.05	.10
$R^{(48,36)}$.03	.05	.02	.04
.
.
.
$R^{(48,47)}$.01	.01	.01	.01

* $\ell = 48$ for the "final" trend-cycle filter and $\ell = 42$ for the final seasonal adjustment filter. However, the values shown for the revision of the seasonal adjustment filters are also calculated for $\ell = 48$ since after $\ell = 42$ the values are final and, thus, do not change.

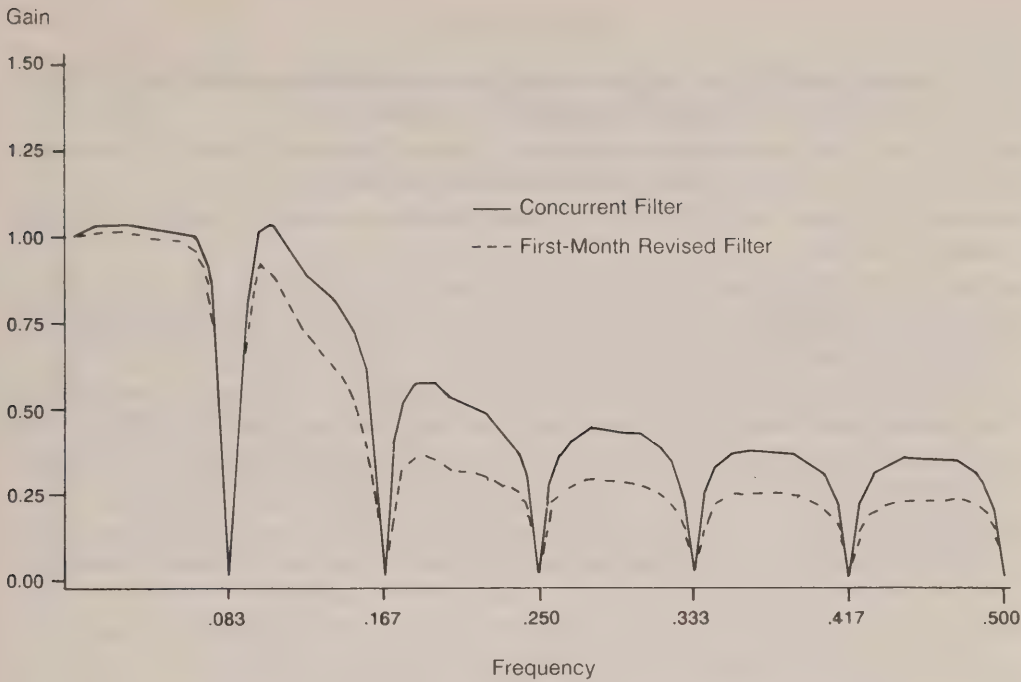


Figure 3a. Gain Functions of the Concurrent and First-Month Revised Filters of X-11-ARIMA with ARIMA Extrapolations ($\theta = .40$, $\Theta = .60$).

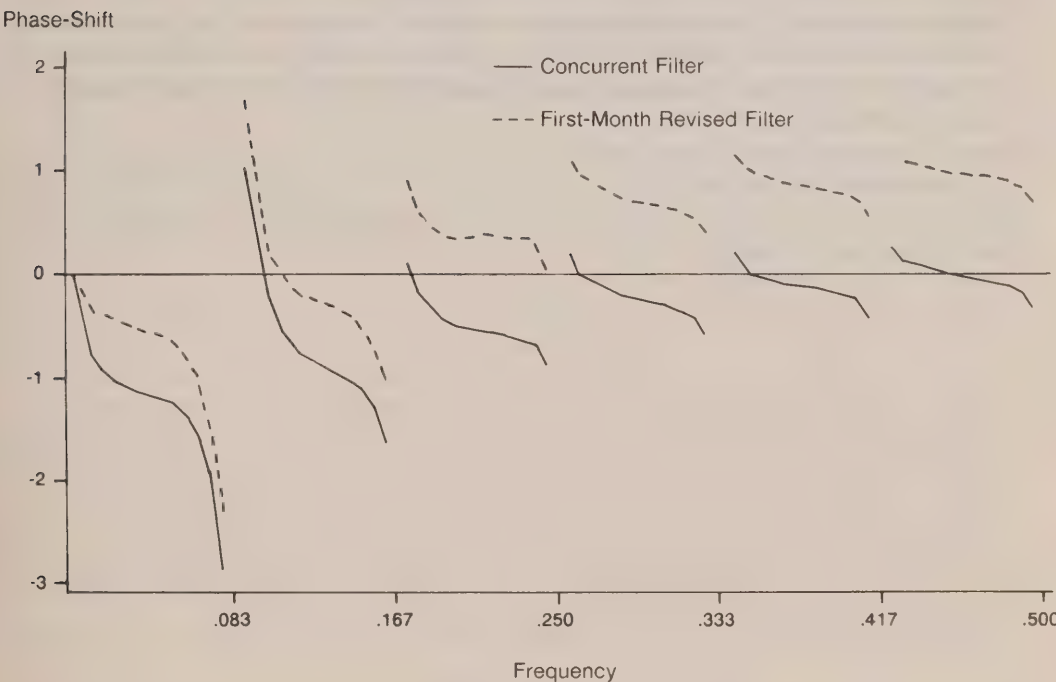


Figure 3b. Phase-Shift Functions of the Concurrent and First-Month Revised Filters of X-11-ARIMA with ARIMA Extrapolations ($\theta = .40$, $\Theta = .60$).

REFERENCES

- BAYER, A., and WILCOX, D. (1981). An evaluation of concurrent seasonal adjustment. Technical Report, Board of Governors of the Federal Reserve System.
- DAGUM, E.B. (1978). *Comparison and Assessment of Seasonal Adjustment Methods for Labour Force Series*. Stock No. 052-003-00603-1, U.S. Government Printing Office.
- DAGUM, E.B. (1980). *The X-11-ARIMA Seasonal Adjustment Method*. Catalogue No. 12-564E, Statistics Canada.
- DAGUM, E.B. (1982a). Revisions of time varying seasonal filters. *Journal of Forecasting*, 1, 173-187.
- DAGUM, E.B. (1982b). The effects of asymmetric filters on seasonal factor revisions. *Journal of the American Statistical Association*, 77, 732-738.
- DAGUM, E.B. (1982c). Revisions of seasonally adjusted data due to filter changes. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 39-45.
- DAGUM, E.B., and MORRY, M. (1984). Basic issues on the seasonal adjustment of the Canadian Consumer Price Index. *Journal of Business and Economic Statistics*, 2, 250-259.
- DAGUM, E.B. (1987). Monthly versus annual revisions of concurrent seasonally adjusted series. In *Time Series and Econometric Modelling*, (Eds. I.B. MacNeill and G.J. Umphrey), New York: D. Reidel, 131-196.
- DAGUM E.B., and LANIEL, N. (1987). Revisions of trend-cycle estimators of moving average seasonal adjustment method. *Journal of Business and Economic Statistics*, (forthcoming).
- KENNY, P., and DURBIN, J. (1982). Local trend estimation and seasonal adjustment of economic time series. *Journal of the Royal Statistical Society, Ser. A*, 145, 1-41.
- MARAVALL, A. (1986). An application of model-based estimation of unobserved components. *International Journal of Forecasting*, 2, 305-318.
- MOORE, G.H., BOX, G.E.P., KAITZ, H.B., STEPHENSON, J.A., and ZELLNER, A. (1981). Seasonal adjustment of the monetary aggregates. In *Report of the Committee of Experts on Seasonal Adjustment Techniques*, Washington: Board of Governors of the Federal Reserve System.
- McKENZIE, S. (1984). Concurrent seasonal adjustment with Census X-11. *Journal of Business and Economic Statistics*, 2, 235-249.
- PIERCE, D.A. (1980). Data revisions with moving average seasonal adjustment procedures. *Journal of Econometrics*, 14, 95-114.
- PIERCE, D., and McKENZIE, S. (1985). On concurrent seasonal adjustment. Special Studies Paper 164, Federal Reserve Board.

On Efficient Estimation of Unemployment Rates from Labour Force Survey Data

S. KUMAR and A.C. SINGH¹

ABSTRACT

The method of minimum $Q^{(T)}$ estimation for complex survey designs proposed by Singh (1985) provides asymptotically efficient estimates of model parameters analogous to Neyman's (1949) min X^2 estimation procedure for simple random samples. The $Q^{(T)}$ can be viewed as a X^2 type statistic for categorical survey data, and min $Q^{(T)}$ estimates provide a robust alternative to Weighted Least Squares estimates, which often display unstable behaviour for complex surveys. In this paper, the min $Q^{(T)}$ method is first described and then illustrated for the problem of estimating parameters of a logit model for survey estimates of unemployment rates which are obtained from the October 1980 Canadian LFS data cross-classified according to age-education covariate categories. It is seen that the trace efficiency of smoothed estimates obtained by Kumar and Rao (1986), who applied the method of pseudo maximum likelihood estimates (pseudo mle) to the same problem can be slightly improved by the min $Q^{(T)}$ method. Interestingly enough, pseudo mle for individual cells behave much the same way as the efficient min $Q^{(T)}$ estimates for the particular LFS example.

KEY WORDS: Pseudo mle; WLS estimator; Min $Q^{(T)}$ estimator; Asymptotic efficiency; Approximate likelihood; Generalized score statistic.

1. INTRODUCTION

Based on October 1980 Labour Force Survey (LFS) data, Kumar and Rao (1984, 1986) proposed and analysed a logistic regression (logit) model for unemployment rates. They used the theory developed by Roberts (1985) and Roberts, Rao and Kumar (1987) who generalized the Rao-Scott method (1981, 1984) of adjusting X^2 for impact of the underlying survey design to test the fit of the logit model. Kumar and Rao considered unemployment rates in various cells (or domains) that had been obtained by cross-classifying the population into a number of age and education categories. The logit model consisted of both linear and quadratic effects for the age variable, with only the linear effect for the education variable. The same LFS data were also analysed by Singh and Kumar (1986) using an alternative method, namely the $Q^{(T)}$ test proposed by Singh (1985). The test $Q^{(T)}$ is a X^2 type test based on a generalized score statistic of principal components. Results obtained by the $Q^{(T)}$ method were found to be in agreement with those arrived at by the adjusted X^2 method.

Whenever a suitable model is determined, it is of interest to find good estimates of model parameters. These, in turn, provide fairly good estimates of true rates for domains. Such estimates (often called "smoothed estimates") are especially useful for domains in which survey estimates lack precision because the number of observations is not sufficient. It may be noted that since smoothed estimates are obtained after a model is found to have a reasonable fit, the bias in the estimates is expected to be negligible. Kumar and Rao (1986) used the method of pseudo mle (pseudo maximum likelihood estimates) under the working form of the likelihood that corresponds to independent binomial samples for estimating parameters

¹ S. Kumar, Senior Methodologist, Social Survey Methods Division, Jean Talon Building, Tunney's Pasture, Statistics Canada, Ottawa, Ontario, K1A 0T6. A.C. Singh, Associate Professor, Department of Mathematics and Statistics, Memorial University of Newfoundland, St. John's, Newfoundland, Canada, A1C 5S7.

of a logit model after an adequate fit had been established for the October 1980 LFS data. They found a considerable gain in efficiency over survey estimates of unemployment rates in the particular LFS example.

Pseudo mle are known to be useful when the likelihood function is not available or when it is difficult to compute due to complexities of the survey design. Under suitable regularity conditions, the pseudo mle provide consistent and asymptotically normal estimates (Imrey, Koch and Stokes 1982). In this paper we consider the problem of finding asymptotically efficient (in a sense to be explained in Section 3) estimates of model parameters and therefore of domain estimates. We describe the $\min Q^{(T)}$ estimator, proposed in Singh (1985), based on the generalized scores approach which can be viewed as analogous to Neyman's $\min X^2$ estimator for simple random samples. It may be noted that the WLS (Weighted Least Squares) approach for complex survey designs (Koch, Freeman and Freeman 1975) also provides asymptotically efficient estimates. However, these estimates are usually unstable for moderate sample sizes due to near singularity of the estimated covariance matrix of survey cell estimates (see Imrey, Koch and Stokes 1982, Fay 1985). The $\min Q^{(T)}$ estimates, on the other hand, are designed to guard against the instability problem mentioned above. It will be seen that the problem of instability can be overcome by the $\min Q^{(T)}$ method by employing a modified version of the estimated covariance matrix in which the relatively very small eigenvalues from its spectral decomposition are trimmed.

The necessary notation along with a brief review of the test $Q^{(T)}$ are presented in Section 2. Next the $\min Q^{(T)}$ estimator and its asymptotic behaviour are described in Section 3. The example using LFS data is given in Section 4 as an illustration. For this numerical example, an interesting finding was that over individual cells, the pseudo mle perform almost at par with efficient $\min Q^{(T)}$ estimates. In terms of an overall measure as given by trace efficiency, pseudo mle are found to be only slightly inferior to $\min Q^{(T)}$ estimates. Finally, Section 5 contains some concluding remarks.

2. THE TEST $Q^{(T)}$: A BRIEF REVIEW

We shall briefly describe the test $Q^{(T)}$ in order to motivate the $\min Q^{(T)}$ method of estimation (for more details, see Singh 1985, Singh and Kumar 1986). Let I denote the number of disjoint domains and v_i denote the parameter of interest for the i -th domain. Consider a model for $v = (v_1, v_2, \dots, v_I)'$ as

$$H_0: h(v) = X\theta \quad (2.1)$$

where X is a known $I \times r$ matrix of full rank r , θ is an r -vector of unknown parameters, and h is a continuously differentiable one-to-one function, for instance, log or logit.

Let \hat{v} denote the I -vector of survey estimates. Assume that under a suitable central limit theorem

$$\hat{v} \sim MVN(v, \Gamma/n) \quad (2.2)$$

where " \sim " means "asymptotically distributed as", n is the total sample size, and Γ is the asymptotic covariance matrix of $\sqrt{n}(\hat{v} - v)$.

Now, choose a small level $\epsilon (>0)$ of dimensionality reduction (eg., .01 or .005 can be taken as working values of ϵ). Find a number T such that with the eigenvalues $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_I$ of the estimated covariance matrix $\hat{\Gamma}$, we have

$$T = \max \left\{ t: t > r \text{ and } \sum_{i=t}^I \hat{\lambda}_i / \sum_{i=1}^I \hat{\lambda}_i \geq \epsilon \right\}. \tag{2.3}$$

The variable T , although random, can be regarded as fixed for our asymptotics. It may be noted that if there are no relatively very small eigenvalues (i.e. if $\hat{\Gamma}$ is not ill-conditioned), then there will usually be no effect of dimensionality reduction for small ϵ and T will coincide with I in those situations.

Consider the problem of testing H_0 against alternatives $K_0: h(v) \neq X\theta$ in the class of tests based on the first T principal components W of \hat{v} . Let the normalized eigenvector corresponding to $\hat{\lambda}_i$ be P_i (it need not be unique) and let M_T denote the $I \times T$ matrix of eigenvectors P_i 's corresponding to the first T largest eigenvalues. Then

$$W = M_T' \hat{v} \sim MVN(\mu, D_T/n), \tag{2.4}$$

where

$$\mu = M_T' v, D_T = \text{diag}(\lambda_1, \dots, \lambda_T).$$

Based on W , the original testing problem concerning an I -dimensional v is reduced to testing a hypothesis about the T -dimensional parameter μ given by

$$H_0': \mu = M_T' h^{-1}(X\theta) \text{ vs } K_0': \mu \neq M_T' h^{-1}(X\theta). \tag{2.5}$$

The test statistic $Q^{(T)}$ can be obtained as a score statistic of principal components by employing the approximate likelihood of θ given by the limiting distribution (2.4) of W for computing the efficient scores (see Cox and Hinkley 1974, p. 321-324). We shall refer to $Q^{(T)}$ as a generalized score test that would reject H_0 for large values of the quadratic form

$$\begin{aligned} Q^{(T)}(\theta^o) &= Y(\theta^o)' \Delta_T Y(\theta^o) - Z_T(\theta^o)' \wedge_T Z_T(\theta^o) \\ &\sim \chi^2_{T-r} \end{aligned} \tag{2.6}$$

where

$$\begin{aligned} Y(\theta^o) &= \hat{v} - v(\theta^o), \Delta_T = n \sum_{i=1}^T (P_i P_i' / \hat{\lambda}_i), \\ Z_T(\theta^o) &= B' \Delta_T Y(\theta^o), B = (\partial v / \partial \theta), \wedge_T = (B' \Delta_T B)^{-1}, \end{aligned}$$

and θ^o is some fixed point in the null parameter space. In computing $Q^{(T)}$, any root n -consistent estimate of θ under H_0 can be substituted for θ^o , such as pseudo mle of θ . Notice that $Q^{(T)}$ of (2.6) is in fact a quadratic form in W but is expressed in \hat{v} for the sake of convenience.

For testing H_0 vs K_0 in the class of tests based on W , the asymptotic optimality of the test $Q^{(T)}$ follows from that of the score statistic. For small $\epsilon > 0$, \hat{v} and W will be close in the sense that principal components provide the best possible way of dimensionality reduction

with a minimum loss of information. Thus $Q^{(T)}$ (for small ϵ) is expected to be robust with respect to the test Q corresponding to no dimensionality reduction. However, Q may be unstable (in the sense of inflated Type I error rate) for finite samples due to possible near singularity of $\hat{\Gamma}$. The test $Q^{(T)}$ is expected to control this problem of instability at the cost of sacrificing some information in the data that gives rise to possibly unreliable components in Q in the directions of eigenvectors that correspond to relatively very small eigenvalues. The loss of information implies that the test $Q^{(T)}$ will lack power for alternatives in directions of (near) singularities. However, this loss of power is offset by the gain in control of Type I error rate. The instability control is further ensured by the fact that, since H_0 is a subset of H'_0 , $Q^{(T)}$ will be a conservative test for H_0 .

A special asymptotically equivalent version of $Q^{(T)}$ (θ^0) which has a simpler expression similar to that of the standard Pearson-Fisher's X^2 , is obtained by replacing θ^0 with an estimator $\tilde{\theta}$ that minimizes the expression $(\hat{v} - v(\theta))' \Delta_T (\hat{v} - v(\theta))$. We then have

$$\begin{aligned} Q^{(T)}(\tilde{\theta}) &= Y(\tilde{\theta})' \Delta_T Y(\tilde{\theta}) \\ &= \sum_{i=1}^T [P'_i (\hat{v} - v(\tilde{\theta}))]^2 / \hat{\lambda}_i \\ &\sim \chi^2_{T-r} \end{aligned} \quad (2.7)$$

Henceforth we assume that, for a given data vector \hat{v} , a model H_0 has been deemed appropriate based on the test $Q^{(T)}$ or some other test such as the adjusted X^2 test. In the next section, we give an asymptotically efficient method of estimating parameters θ under H_0 , using the statistic $Q^{(T)}$. The θ estimates in turn provide a set of smoothed estimates of v corresponding to survey estimates \hat{v} .

3. THE MIN $Q^{(T)}$ ESTIMATOR

Consider the approximate likelihood for the mean μ of the first T principal components W of \hat{v} , given earlier by (2.4). Suppose the model $H_0: h(v) = X\theta$ is accepted. Then, the kernel function $K(\theta)$ of the approximate likelihood for $\mu(\theta)$ is given by

$$\begin{aligned} K(\theta) &= (W - \mu(\theta))' D_T^{-1} (W - \mu(\theta)) \\ &= (\hat{v} - v(\theta))' \Delta_T (\hat{v} - v(\theta)) \end{aligned} \quad (3.1)$$

The value $\tilde{\theta}$ that minimizes $K(\theta)$ corresponds to the mle of θ for the approximate likelihood of μ under H_0 . The estimator $\tilde{\theta}$ will be asymptotically efficient (or best asymptotically normal (BAN) in the sense of Neyman, 1949), in a restricted class, namely in the class of estimates based on W . Following the min X^2 estimator of Neyman (1949), the estimator $\tilde{\theta}$ was termed min $Q^{(T)}$ estimator in Singh (1985). Notice that the estimator $\tilde{\theta}$ depends on the level ϵ of dimensionality reduction via Δ_T . Thus $\tilde{\theta}$ varies if ϵ does.

The smoothed estimates of v under H_0 based on W can be obtained as follows. Find $\tilde{\theta}$ which minimizes $K(\theta)$, i.e. $\tilde{\theta}$ is the solution of r equations

$$B' \Delta_T (\hat{v} - v(\theta)) = 0 \quad (3.2)$$

where both $B(=\partial v/\partial \theta)$ and v involve θ . An iterative procedure such as Newton-Raphson can be used to solve (3.2). Weighted least squares (WLS) estimates or pseudo mle can be used as possible initial choices for θ . We can then compute the $Q^{(T)}$ estimator of v as

$$\tilde{v} = h^{-1}(X\tilde{\theta}). \tag{3.3}$$

The asymptotic behaviours of $\tilde{\theta}$ and \tilde{v} are given by the following proposition.

Proposition 3.1 As before, let Λ_T denote $(B'\Delta_TB)^{-1}$. We have

(a) $\tilde{\theta} - \theta \approx \Lambda_TB'\Delta_T(\hat{v} - v(\theta)) \rightsquigarrow MVN(0, \Lambda_T)$

(b) $\tilde{v} - v \approx B\Lambda_TB'\Delta_T(\hat{v} - v(\theta)) \rightsquigarrow MVN(0, B\Lambda_TB')$

(3.4)

where “ \approx ” indicates that the difference between the two sides is negligible in probability.

The proof follows from the application of the δ -method to the functions $B'\Delta_T(\hat{v} - v(\theta))$ and $\tilde{v} - v(\theta)$, which gives

$$\begin{aligned} B'\Delta_T(\hat{v} - v(\theta)) - (B'\Delta_TB)(\tilde{\theta} - \theta) &= o_p(1), \\ \tilde{v} - v(\theta) - B(\tilde{\theta} - \theta) &= o_p(1). \end{aligned}$$

From the above proposition it follows that the asymptotic covariance matrix of the $\min Q^{(T)}$ estimator $\tilde{\theta}$ is the inverse of the information matrix $B'\Delta_TB$ for θ , which was obtained from the approximate likelihood of θ as given by (2.4). It can then be seen that in the absence of dimensionality reduction, the estimator $\tilde{\theta}$ will be asymptotically equivalent to the WLS estimator of Koch, Freeman and Freeman (1975). As mentioned in the Introduction, the WLS estimator generally shows unstable finite sample behaviour because of the inefficient estimation of Γ . In contrast, the estimator $\tilde{\theta}$ for a given $\epsilon > 0$ is expected to show stable finite sample behaviour in the sense that it can be approximated well by its asymptotic behaviour. This is achieved at the cost of compromising the asymptotic optimality of $\tilde{\theta}$ by restricting it to a smaller class, namely the class of estimates based on the first T principal components W . The WLS estimator, on the other hand, is asymptotically optimal in a wider class, namely the class of estimates based on the full data vector \hat{v} . If, for a small ϵ , the $Q^{(T)}$ test statistic indicates insignificance for H_0 , then the corresponding $\min Q^{(T)}$ estimator \hat{v} will likely provide a robust alternative to the WLS estimator.

4. MIN $Q^{(T)}$ ESTIMATES OF UNEMPLOYMENT RATES

The Canadian labour force survey (LFS) data for October 1980 was analysed by Kumar and Rao (1984, 1986) and Roberts, Rao and Kumar (1987). Both sets of authors applied the extension of the Rao-Scott adjusted X^2 method to the case of logistic regression. They showed that the logit model given below provided an adequate fit to the survey estimates of employment rates ($v_{j\ell}$) for the table of 60 cells cross-classified by age (10 categories) and education (6 categories). The model is

$$\log \frac{v_{j\ell}}{1 - v_{j\ell}} = \beta_0 + \beta_1 A_j + \beta_2 A_j^2 + \beta_3 E_\ell \tag{4.1}$$

where A_j represents the midpoint $12 + 5j$ for j -th age group ($j = 1, \dots, 10$), and E_ℓ ($\ell = 1, \dots, 6$) represents the median years of schooling with values 7, 10, 12, 13, 14 and 16.

The model (4.1) can be expressed in the notation of Section 2 by numbering the sixty cells lexicographically. Thus, (4.1) can be rewritten as $h(v) = X\theta$, where v is the vector of employment rates, h is the logit function, X is a 60×4 matrix whose i -th row is $(1, A_i, A_i^2, E_i)$, and θ is $(\beta_0, \beta_1, \beta_2, \beta_3)'$. We also have

$$H = (\partial h / \partial v) = D_v^{-1} D_{1-v}^{-1}, B = H^{-1} X, \quad (4.2)$$

where D_v and D_{1-v} are diagonal matrices with diagonal elements given by the subscripts.

The pseudo mle of θ for the model (4.1) were obtained by Kumar and Rao (1984) under the pseudo product-binomial likelihood as

$$\bar{\theta} = (-3.10, 0.211, -0.00218, 0.1509)'. \quad (4.3)$$

They also computed Rao-Scott's first order adjusted X^2 (G_c^2 in their notation) as 55.3, which shows acceptance of the model (2.1) when referred to the χ_{56}^2 distribution.

The $Q^{(T)}$ method was applied for testing (4.1) (see Singh 1985, and Singh and Kumar 1986) also resulting in the acceptance of the model (4.1). For $\epsilon = .01$, T turns out to be 51 using the estimated covariance matrix $\hat{\Gamma}$ as obtained by Kumar and Rao (1984). Now using the pseudo mle $\bar{\theta}$, we have

$$Q^{(51)}(\bar{\theta}) = 58.665 - 4.454 = 54.211 \quad (4.4)$$

When $\epsilon = .005$, T is found to be 54, and

$$Q^{(54)}(\bar{\theta}) = 67.774 - 2.343 = 65.431 \quad (4.5)$$

When $\epsilon = 0$, $T = 58$ because two cells had zero observed unemployment rates. In this case,

$$Q^{(58)}(\bar{\theta}) = 87.302 - 0.812 = 86.49 \quad (4.6)$$

By referring $Q^{(51)}$ to the χ_{47}^2 distribution, $Q^{(54)}$ to a χ_{50}^2 and $Q^{(58)}$ to a χ_{54}^2 distribution, it is clear that both $Q^{(51)}$ and $Q^{(54)}$ accept (4.1) while $Q^{(58)}$ does not. An instability check can be performed by considering the difference $Q^{(58)} - Q^{(T)}$ for $T = 51, 54$, which can be seen to be highly significant when referred to the χ_{58-T}^2 distribution. These indicate presence of the instability problem in the Q -test statistic that corresponds to no dimensionality reduction. It is clear that WLS test would also have an instability problem due to the difficulty involved in inverting the matrix $\hat{\Gamma}$ which is singular. Thus, min $Q^{(T)}$ method would be preferable to min Q or WLS methods. In the interests of reducing loss of information, the method with the largest value of T is recommended, providing of course that the corresponding $Q^{(T)}$ shows insignificance for the model.

We shall now compute asymptotically efficient estimates. Neither min Q nor WLS estimates were computed because $\hat{\Gamma}$ was singular. The min $Q^{(T)}$ estimates $\tilde{\theta}$ were computed for $\epsilon = .005$ and $\epsilon = .01$ by using the Newton-Raphson iterative procedure and θ as the initial estimate of θ for solving (3.2). The values of $\tilde{\theta}_T$ and $Q^{(T)}(\tilde{\theta})$ (in this case the negative term in (2.6) drops out) for $\epsilon = .005$, $T = 54$ were obtained as

$$\tilde{\theta}_{54} = (-2.7112, 0.1944, -0.00196, 0.1432)', \text{ and}$$

$$Q^{(54)}(\tilde{\theta}_{54}) = 63.4737 \quad (4.7)$$

For $\epsilon = .01, T = 51$, we have

$$\tilde{\theta}_{51} = (-2.6739, 0.19702, -0.00202, 0.1364)', \text{ and}$$
$$Q^{(51)}(\tilde{\theta}_{51}) = 55.2518.$$

(4.8)

Conclusions based on the statistic $Q^{(T)}(\tilde{\theta})$ for both $T = 54$ and 51 agree with those obtained from $Q^{(T)}(\bar{\theta})$.

Table 1 gives efficiencies relative to survey estimates of unemployment rates $1 - v$ for all cells (excepts two with zero observed unemployment rates) corresponding to the three smoothed estimates. The three smoothed estimates are the pseudo mle, $\min Q^{(51)}$, and $\min Q^{(54)}$. The pseudo mle variances are taken from Kumar and Rao (1986), while those for $\min Q^{(54)}$ estimates are obtained from the diagonal elements of $B \wedge_T B'$ of (3.4). As noted by Kumar and Rao (1986) for pseudo mle, smoothed estimates based on $\min Q^{(T)}$ also lead to considerable efficiency gains over survey estimates. The relative trace efficiency of smoothed estimates over survey estimates is 17.9 for pseudo mle, 18.95 for $\min Q^{(51)}$ and 19.88 for $\min Q^{(54)}$ estimates. Thus the $\min Q^{(T)}$ estimators provide a slight improvement in the

Table 1
Efficiencies of Smoothed Estimates of Unemployment rates
relative to Survey Estimates^a

Cell Number	Min $Q^{(51)}$	Min $Q^{(54)}$	Pseudo mle	Cell Number	Min $Q^{(51)}$	Min $Q^{(54)}$	Pseudo mle
1	5.87	5.74	5.44	31	9.01	9.32	8.65
2	3.62	3.62	3.28	32	8.76	9.46	10.68
3	3.45	3.55	3.12	33	36.93	42.93	51.59
4	52.45	51.65	43.46	34	51.55	60.23	81.12
5	104.77	114.30	96.21	35	69.76	79.93	98.37
7	5.33	5.14	4.38	36	9.17	11.01	15.07
8	9.36	9.53	8.09	37	3.48	3.01	3.45
9	6.85	7.16	6.70	38	13.74	15.91	18.00
10	25.65	28.40	26.31	39	66.87	80.98	97.30
11	13.34	14.13	17.73	40	154.81	187.73	221.50
12	27.74	30.85	30.85	41	49.14	67.56	80.61
13	8.64	8.84	7.15	42	17.32	21.73	24.98
14	13.84	13.84	12.37	43	8.57	9.28	8.49
15	8.20	8.49	9.47	44	27.42	31.65	30.74
16	23.14	24.09	27.75	45	58.55	70.67	75.72
17	18.20	18.20	21.49	46	94.11	114.13	121.49
18	9.87	11.14	12.51	47	82.12	112.65	108.52
19	15.87	16.03	13.66	48	26.54	39.41	41.22
20	11.44	11.98	12.56	49	4.95	5.37	4.41
21	12.39	12.39	15.53	50	12.11	14.10	11.17
22	24.83	24.83	32.02	51	6.75	8.61	7.50
23	16.43	18.16	21.55	52	8.83	11.45	9.90
24	6.98	7.83	10.06	53	52.64	71.49	61.14
25	7.49	7.74	6.99	55	3.59	3.93	3.03
26	10.33	11.33	12.32	56	7.33	8.96	8.23
27	6.47	7.18	8.69	57	23.50	29.83	22.11
28	125.81	140.57	172.91	58	221.23	294.59	208.77
29	33.88	38.13	52.00	59	6.45	8.82	6.62
30	14.89	15.24	20.43	60	38.90	52.84	41.96

^a Cells 6 and 54 are omitted due to zero observed unemployment rates.

efficiency of smoothed estimates compared to pseudo mle. With regard to performance over individual cells Table 1 indicates that the pseudo mle behave very well as compared to efficient $Q^{(T)}$ estimates for the example under consideration.

5. CONCLUDING REMARKS

For computing pseudo mle, the working form of the likelihood function corresponds to simple random samples (i.e. multinomial or product-multinomial sampling). The pseudo mle do provide consistent estimates of model parameters without requiring an estimate of the covariance matrix Γ . However, the pseudo mle are not asymptotically efficient for complex survey data. By contrast, the $\min Q^{(T)}$ estimates are asymptotically efficient with respect to the class of estimates based on W (the first T principal components of the vector \hat{v} of survey estimates). For investigating the relative performance of pseudo mle and $\min Q^{(T)}$, it would be desirable to perform a simulation study for efficiency comparisons. The $\min Q^{(T)}$ estimates do take into account of the underlying complex design by employing an appropriate $\hat{\Gamma}$. If $\hat{\Gamma}$ is not ill-conditioned, i.e. it has no relatively very small eigenvalues, then there is no instability problem with the well known WLS estimates which are of course asymptotically efficient. In this case, it will usually turn out that there is no dimensionality reduction for small ϵ , that T will coincide with I and that there will be no loss in efficiency of $\min Q^{(T)}$ estimates in comparison with WLS estimates. However, given the instability problem common with cross-classified categorical survey data, the $\min Q^{(T)}$ estimates are expected to provide a robust alternative to WLS estimates.

ACKNOWLEDGEMENT

The second author's research was supported by Statistics Canada and the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

- COX, D.R., and HINKLEY, D.W. (1974). *Theoretical Statistics*. London: Chapman and Hall
- FAY, R.E. (1985). A jackknifed chi-squared test for complex samples. *Journal of the American Statistical Association*, 80, 148-157.
- IMREY, P.B., KOCH, G.G., and STOKES, M.E. (1982). Categorical data analysis: Some reflections on the log-linear model and logistic regression. Part II: Data analysis. *International Statistical Review*, 50, 35-63.
- KOCH, G.G., FREEMAN, D.H. Jr., and FREEMAN, J.L. (1975). Strategies in the multivariate analysis of data from complex surveys. *International Statistical Review*, 43, 59-78.
- KUMAR, S., and RAO, J.N.K. (1984). Logistic regression analysis of Labour Force Survey Data. *Survey Methodology*, 10, 62-81.
- KUMAR, S., and RAO, J.N.K. (1986). On smoothed estimates of unemployment rates from labour force survey data. In *Small Area Statistics: An International Symposium '85* (Eds. R. Platek, and M.P. Singh), Ottawa: Carleton University.
- NEYMAN, J. (1949). Contribution to the Theory of the X^2 test. In *Proceedings of the First Berkeley Symposium on Mathematical Statistics and Probability* (Ed. J. Neyman), Berkeley: University of California Press, 230-273.

- RAO, J.N.K., and SCOTT, A.J. (1981). The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence in two way tables. *Journal of the American Statistical Association*, 76, 221-230.
- RAO, J.N.K., and SCOTT, A.J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *Annals of Statistics*, 12, 46-60.
- ROBERTS, G.R. (1985). *Contributions to chi-squared tests with survey data*. Ph.D. dissertation, Carleton University, Ottawa.
- ROBERTS, G., RAO, J.N.K., and KUMAR, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, 74, 1-12.
- SINGH, A.C. (1985). On optimal asymptotic tests for analysis of categorical data from sample surveys. Working Paper, Social Survey Methods Division, Statistics Canada.
- SINGH, A.C., and KUMAR, S. (1986). Categorical data analysis for complex surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, (forthcoming).

A Sampling Procedure with Inclusion Probabilities Proportional to Size

A. DEY and A.K. SRIVASTAVA¹

ABSTRACT

A new unequal probability sampling scheme for selecting n (> 2) units without replacement from a finite population is proposed. This scheme ensures that the inclusion probabilities are proportional to sizes. It has the advantage of simplicity in selection and estimation and also provides a non-negative variance estimator. The variance of the Horvitz-Thompson (H-T) estimator under the proposed scheme is shown to be smaller than that of the customary estimator in probability proportional to size sampling with replacement. The proposed scheme also compares favourably with the without replacement scheme suggested by Sampford (1967) in an empirical study on a few natural populations.

KEY WORDS: Unequal probability sampling; Horvitz-Thompson estimator.

1. INTRODUCTION

In unequal probability sampling of n units without replacement from a finite population containing N units, if π_i denotes the inclusion probability of the i -th unit in the sample $i = 1, 2, \dots, N$, the Horvitz and Thompson (1952) estimator (H-T estimator) of Y , the population total of the study variable y , is given by

$$\hat{Y} = \sum_{i \in s} (y_i / \pi_i), \quad (1.1)$$

where y_i is the y -value for the i -th unit and the summation extends over the units included in the sample. The variance of \hat{Y} is

$$Var(\hat{Y}) = \sum_{i=1}^N \sum_{j>i}^N (\pi_i \pi_j - \pi_{ij}) (y_i / \pi_i - y_j / \pi_j)^2, \quad (1.2)$$

where π_{ij} denotes the joint inclusion probability of the i -th and j -th units in the sample ($i \neq j, i, j = 1, 2, \dots, N$).

Considerable reduction in the variance of \hat{Y} can be expected if the sampling scheme ensures that π_i are proportional to a given measure of size, say, x_i for $i = 1, 2, \dots, N$, where it is assumed that x_i are nearly proportional to y_i . Sampling schemes in which π_i are proportional to x_i are termed Inclusion Probability Proportional to Size (IPPS) schemes. For a comprehensive account of unequal probability sampling procedures, including IPPS sampling schemes, the reader is referred to the monograph of Brewer and Hanif (1983).

Some desirable properties of an unequal probability scheme without replacement in general, and IPPS schemes in particular, are simplicity in selection and estimation, availability of a non-negative variance estimator, and better efficiency than with the probability proportional to size (PPS) with replacement strategy. Unfortunately, for sample size greater than two, not many of the available procedures meet these requirements fully.

¹ A. Dey and A.K. Srivastava, Indian Agricultural Statistics Research Institute, Library Avenue, New Delhi 110012, India.

In this paper, an IPPS sampling scheme is suggested for arbitrary sample sizes, $n > 2$. The procedure is rather simple both in sample selection and at the estimation stage since compact expressions for π_{ij} are available. It has also been possible to provide a positive estimator of variance of the H - T estimator of Y . The performance of the H - T estimator under the proposed scheme is compared with the PPS with replacement strategy and a simple sufficient condition is derived under which the performance of the former strategy is superior to that of the latter. An empirical study on a few natural populations indicates that the proposed scheme compares favourably with that suggested by Sampford (1967).

2. THE SAMPLING PROCEDURE

Consider a population of N units with y as the study variable and x , an auxiliary variable, as the size. It is assumed that x -values are known for all the population units. A sample of size $n (> 2)$ is to be selected. To start with, it is assumed that n is even.

Divide the population into $m (> n/2)$ groups such that the i -th group contains $N_i (> 2)$ units ($i = 1, 2, \dots, m$) and, for each group,

$$X_i/X > (n - 2)/[n(m - 1)], \quad (2.1)$$

where

$$X_i = \sum_{u=1}^{N_i} x_{iu},$$

x_{iu} is the value of x for the u -th unit in the i -th group and $X = X_1 + X_2 + \dots + X_m$.

Equation (2.1) is satisfied if the X_i ($i = 1, 2, \dots, m$) are made nearly equal. It has been seen in actual populations, considered by Rao and Bayless (1969) and others, that this condition is satisfied for quite a few values of m if the groups are so formed that their sizes, X_i , are nearly equal. Rao and Lanke (1984) suggested a grouping procedure in which N units are grouped into R groups such that group totals, X_i , are nearly equal and group sizes are either $[N/R]$ or $[N/R] + 1$, where $[x]$ is the largest integer contained in x . For the formation of groups, the Rao-Lanke procedure may also be tried.

Having formed the m groups, the suggested sampling procedure consists of the following steps:

Step 1. Select $n/2$ groups out of the m groups using Midzuno's (1951) sampling procedure with probabilities $\{P'_i\}$, that is, select one group with probability

$$P'_i = [n(m - 1)P_i - (n - 2)]/(2m - n), \text{ with } P_i = X_i/X,$$

and the remaining $(n/2) - 1$ groups with equal probabilities without replacement.

Step 2. From each of the selected groups, select two units by any IPPS procedure, say by Durbin's (1967) procedure, that is, in the i -th selected group ($i = 1, 2, \dots, n/2$) select one unit with probability

$$p_{iu|i} = x_{iu}/X_i,$$

and the second unit with revised probability

$$p_{i_u|i_v} = x_{i_v} [1/(X_i - 2x_{i_v}) + 1/(X_i - 2x_{i_u})] / D_i,$$

where

$$D_i = [1 + \sum_{u=1}^{N_i} x_{i_u} / (X_i - 2x_{i_u})].$$

For this sampling procedure, the inclusion probability for the i_u -th unit is evidently given by

$$\pi_{i_u} = n p_{i_u} \tag{2.2}$$

where

$$p_{i_u} = x_{i_u} / X.$$

Also, the joint inclusion probabilities for a pair of units are given by

$$\pi_{i_u i_v} = \frac{n p_{i_u} p_{i_v} (P_i - p_{i_u} - p_{i_v})}{D_i (P_i - 2 p_{i_u}) (P_i - 2 p_{i_v})} \tag{2.3}$$

and

$$\pi_{i_u j_v} = \frac{n (n - 2) p_{i_u} p_{j_v}}{(m - 1) (m - 2) P_i P_j} [(m - 1) (P_i + P_j) - 1], \tag{2.4}$$

$$i \neq j, i, j = 1, 2, \dots, m.$$

Thus we see that the proposed scheme is indeed an IPPS scheme.

As mentioned earlier, at step 2 of the proposed procedure, any IPPS scheme for selecting two units can be used. Since the procedure of Durbin (1967), which is equivalent to those of Rao (1963) and Brewer (1963), generally performs well, it has been adopted at step 2.

3. A VARIANCE ESTIMATOR

Two well-known unbiased estimators of $Var(\hat{Y})$ are due to Horvitz and Thompson (1952) and Yates and Grundy (1953). Both these estimators, however, suffer from the drawback that they sometimes assume negative values. In this section, a positive estimator of variance is proposed that utilizes the two-stage nature of the proposed sampling scheme.

Using a result due to Des Raj (1966), an unbiased estimator of $Var(\hat{Y})$ is given by

$$\begin{aligned} \hat{V}(\hat{Y}) = & \sum_{i=1}^{n/2} \pi_i^{-1} \sum_{u < v} \sum \left[\frac{\pi_{i_u|i} \pi_{i_v|i}}{\pi_{i_u i_v|i}} - 1 \right] \left[\frac{y_{i_u}}{\pi_{i_u|i}} - \frac{y_{i_v}}{\pi_{i_v|i}} \right]^2 \\ & + \sum_{i < j}^{n/2} \sum_{i < j}^{n/2} \left(\frac{\pi_i \pi_j}{\pi_{ij}} - 1 \right) \left[\frac{\hat{Y}_i}{\pi_i} - \frac{\hat{Y}_j}{\pi_j} \right]^2, \end{aligned} \tag{3.1}$$

where

$$\pi_i = n P_i / 2,$$

$$\pi_{ij} = \frac{n(n-2)}{4(m-2)} \{ (P_i + P_j) - 1/(m-1) \},$$

$$\pi_{i_u|i} = 2p_{i_u}/P_i,$$

$$\pi_{i_u i_v|i} = \frac{2p_{i_u}p_{i_v}(P_i - p_{i_u} - p_{i_v})}{D_i P_i (P_i - 2p_{i_u})(P_i - 2p_{i_v})},$$

and

$$\hat{Y}_i = \sum_{u=1}^2 y_{i_u} / \pi_{i_u|i}, \quad (3.2)$$

y_{i_u} being the y -value of the u -th unit in the i -th group.

The two terms in the right side of (3.1) correspond to the Yates-Grundy variance estimator in Durbin's and Midzuno's procedures. Since under these two sampling procedures the Yates-Grundy estimator of variance is always positive, it follows that the variance estimator given by (3.1) is also positive. However, the estimator in (3.1) is neither the Horvitz-Thompson nor the Yates-Grundy variance estimator.

4. COMPARISON WITH PPS WITH REPLACEMENT STRATEGY

In this section, we compare the efficiencies of the following two strategies:

Strategy 1. The proposed sampling scheme in conjunction with the Horvitz-Thompson estimator.

Strategy 2. PPS sampling with replacement in conjunction with the customary estimator.

Strategy 1 is more efficient than Strategy 2 if and only if

$$\begin{aligned} & \sum_{i=1}^m \sum_{u \neq v}^{N_i} \pi_{i_u i_v} (y_{i_u}/p_{i_u} - Y)(y_{i_v}/p_{i_v} - Y) \\ & + \sum_{i \neq j}^m \sum_{u=1}^{N_i} \sum_{v=1}^{N_j} \pi_{i_u j_v} (y_{i_u}/p_{i_u} - Y)(y_{j_v}/p_{j_v} - Y) < 0. \end{aligned} \quad (4.1)$$

After some lengthy but routine algebra, the inequality (4.1) boils down to

$$\begin{aligned} & - \sum_{i=1}^n (n/D_i) \sum_{u=1}^{N_i} (y_{i_u} - Y_i p_{i_u}/P_i)^2 / (P_i - 2p_{i_u}) \\ & - n(n-2) \left[\sum_{i=1}^m (Y_i/P_i - Y) \right]^2 / [(m-2)(m-1)] \\ & - n(m-2)^{-1} \sum_{i=1}^m [\{ (2n-m-2)P_i - (n-2)(m-1)^{-1} \} (Y_i/P_i - Y)^2] < 0, \end{aligned} \quad (4.2)$$

where

$$Y_i = \sum_u y_{iu}.$$

Obviously, (4.2) holds if

- (i) $(2n - m - 2) > 0$, and
 - (ii) $P_i > (n - 2) / [(m - 1)(2n - m - 2)]$.
- (4.3)

Also, since we are using Midzuno's procedure at the first stage with revised probabilities $\{P'_i\}$, each P_i must satisfy (2.1), that is, each P_i must satisfy

$$P_i > (n - 2) / [n(m - 1)].$$

Thus, (4.2) holds if

$$m \leq (n - 2). \tag{4.4}$$

It appears, therefore, that for Strategy 1 to be superior to Strategy 2, m should be chosen such that

$$n/2 < m \leq (n - 2). \tag{4.5}$$

However, it is clear that (4.4) is merely a sufficient condition and is not necessary. For $n > 6$, condition (4.5) offers a somewhat wide choice for the value of m , while for $n = 6$, (4.5) implies that $m = 3$. For $n = 4$, (4.5) does not lead to a feasible value of m . Therefore, for $n = 4$, an investigation into the performance of Strategy 1 has been taken up for various values of m , not constrained by (4.5), on certain natural populations. A description of the populations appears in Table 1. Table 2 presents the relative efficiency of Strategy 1 compared to Strategy 2 for the populations in Table 1. The performance of the H-T estimator under Sampford's (1967) scheme (called Strategy 3) is also compared with that of Strategy 2.

It can be observed from Table 2 that the performance of the proposed strategy (Strategy 1) compares favourably with that of Sampford (Strategy 3) for most of the populations. Of course, both strategies are superior to Strategy 2.

To achieve the relative efficiency of Strategy 1, the units were grouped in an ad-hoc manner, ensuring only that requirement (2.1) was satisfied. The procedure of Rao and Lanke (1984) was also attempted in forming the groups. However, the Rao-Lanke procedure did not always result in a high efficiency. Further investigations are necessary to decide the 'best' choice of groups. For certain populations, suitable groups satisfying (2.1) could not be formed for higher values of m , and thus, for these cases, the relative efficiencies are not reported in Table 2.

In conclusion, a brief comment on cases in which the desired sample size, n , is odd is in order. An IPPS sample for odd n may be obtained by selecting $(n + 1)$ units by the suggested procedure and then randomly discarding one unit. The expressions for π_i and π_{ij} under this procedure are straightforward. Obviously, when one of the sample units out of $(n + 1)$ is discarded at random, the resulting sample consists of two units from each of the $(n - 1)/2$ groups and just one unit from one of the groups. An unbiased and positive estimator of $Var(\bar{Y})$ can be obtained, analogous to (3.1), on the basis of the $(n - 1)/2$ groups, each containing two units in the sample.

Table 1
Description of the Populations

Pop. Number	Source	N	y	x
1.	Des Raj (1965)	20	Number of households	Eye-estimated number of households
2.	Rao (1963)	14	Corn acreage in 1960	Corn acreage in 1958
3.	Cochran (1963, p. 204)	10	Weight of peaches	Eye-estimated weight of peaches
4.	Hanurav (1967)	20	Population in 1967	Population in 1957
5.	Hanurav (1967)	19	Population in 1967	Population in 1957
6.	Hanurav (1967)	16	Population in 1967	Population in 1957
7.	Hanurav (1967)	17	Population in 1967	Population in 1957
8.	Cochran (1963, p. 325)	10	Number of persons per block	Number of rooms per block
9.	Cochran (1963, p. 156, cities 1-16)	16	Population in 1930	Population in 1920
10.	Cochran (1963, p. 156, cities 33-49)	17	Population in 1930	Population in 1920
11.	Sampford (1962, p. 61)	35	Oats acreage in 1957	Oats acreage in 1947
12.	Sukhatme and Sukhatme (1970, p. 256, circles 1-20)	20	Wheat acreage	Number of villages
13.	Sukhatme and Sukhatme (1970, p. 256, circles 21-40)	20	Wheat acreage	Number of villages
14.	Yates (1960, p. 163)	20	Volume of timber	Eye-estimated volume of timber

Table 2
Percent Relative Efficiencies of
Strategies 1 and 3 over Strategy 2 for the
Populations in Table 1 ($n = 4$)

Pop. Number	Strategy 1				Strategy 3
	$m = 3$	4	5	6	
1.	130.1	118.7	120.8	124.5	127.8
2.	132.6	130.2	—	—	127.1
3.	149.1	—	—	—	147.9
4.	120.7	120.6	122.7	129.7	117.8
5.	129.1	138.7	158.7	—	125.1
6.	158.0	173.1	—	—	139.5
7.	151.9	144.8	169.2	—	131.9
8.	168.5	—	—	—	145.5
9.	118.3	116.3	—	—	109.5
10.	126.6	—	—	—	112.2
11.	113.8	116.2	135.6	129.9	113.8
12.	117.4	128.0	119.0	—	119.3
13.	122.2	120.6	—	—	119.7
14.	124.8	123.1	115.4	113.2	116.3

ACKNOWLEDGEMENTS

The authors would like to thank the referee for making many useful suggestions on the first draft.

REFERENCES

BREWER, K.R.W. (1963). A model of systematic sampling with unequal probabilities. *Australian Journal of Statistics*, 5, 5-13.

BREWER, K.R.W. and HANIF, M. (1983). *Sampling with Unequal Probabilities*. Lecture Notes in Statistics, No. 15. New York: Springer-Verlag.

COCHRAN, W.G. (1963). *Sampling Techniques*, (2nd. ed.). New York: John Wiley.

DES RAJ (1965). Variance estimation in randomized systematic sampling with probability proportional to size. *Journal of the American Statistical Association*, 60, 278-284.

DES RAJ (1966). Some remarks on a simple procedure of sampling without replacement. *Journal of the American Statistical Association*, 61, 391-396.

DURBIN, J. (1967). Design of multi-stage surveys for the estimation of sampling errors. *Journal of the Royal Statistical Society*, Ser. C, 16, 152-164.

HANURAV, T. (1967). Optimum utilization of auxiliary information: π ps sampling of two units from a stratum. *Journal of the Royal Statistical Society*, Ser. B, 29, 379-391.

HORVITZ, D.G. and THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663- 685.

MIDZUNO, H. (1951). On the sampling system with probability proportionate to sum of sizes. *Annals of the Institute of Statistical Mathematics*, 2, 99-108.

RAO, J.N.K. (1963). On three procedures of unequal probability sampling without replacement. *Journal of the American Statistical Association*, 58, 202-215.

- RAO, J.N.K. and BAYLESS, D.L. (1969). An empirical study of the stabilities of estimators and variance estimators in unequal probability sampling of two units per stratum. *Journal of the American Statistical Association*, 64, 540-559.
- RAO, J.N.K. and LANKE, J. (1984). Simplified unbiased variance estimation for multistage designs. *Biometrika*, 71, 387-395.
- SAMPFORD, M.R. (1962). *An Introduction to Sampling Theory*. Edinburgh: Oliver and Boyd.
- SAMPFORD, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54, 499-513.
- SUKHATME, P.V. and SUKHATME, B.V. (1970). *Sampling Theory of Surveys with Applications*, (2nd. ed.). Ames, Iowa: Iowa State University Press.
- YATES, F. (1960). *Sampling Methods for Censuses and Surveys*, (3rd. ed.). London: Griffin.
- YATES, F. and GRUNDY, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society, Ser. B*, 15, 253-261.

Sample Design for the Health and Activity Limitation Survey

D. DOLSON, K. McCLEAN, J.-P. MORIN, and A. THÉBERGE¹

ABSTRACT

The Health and Activity Limitation Survey is part of the program to establish a data base on the disabled population in Canada. The sample design used for the part of the survey covering the population not living in institutions is described. In addition, the methods used to determine the sizes of the samples and to select the samples are presented.

KEY WORDS: Disability; Stratified sampling; Two-stage sampling; Optimum allocation; Sampling without replacement.

1. INTRODUCTION

As part of the program to obtain more information about Canada's disabled population, the Health and Activity Limitation Survey (HALS) was conducted in the fall of 1986. It is designed to obtain information concerning the nature of the problems experienced by that population and, in general, their daily activities (at home, at work, at school, during travel, and so on). The survey is divided into two parts: one covers the population living in institutions and the other, which is the subject of this article, covers the non-institutional population.

Canada has been divided into 238 subprovincial areas (SPAs). All Quebec and Ontario municipalities with more than 125,000 residents and all municipalities in the other provinces with more than 75,000 residents are included as SPA's. The other areas are made up of groups of census subdivisions respecting geographical contiguity and the provincial boundaries. The number of these areas in each province is proportional to the square root of the population, minus the previously defined municipalities. One of the main objectives of the survey is to generate statistics on the disabled population at the SPA level so that the population's various needs can be analysed in detail. In addition, estimates will be produced for three age groups – namely, children (under 15 years of age), adults (15 to 64 years of age) and seniors (65 years of age and older).

The data was collected in two stages. The first stage involved a multipart question (question 20) included on form 2B of the 1986 Canadian Census of Population. This question asked about the respondents' limitations in various types of activities and their own assessments of their conditions. A copy of question 20 is given in the Appendix. The second stage was implemented some time after the census. It involves a screening questionnaire and follow-up to collect information on the problems and activities of disabled respondents.

The main purpose of the first stage is to separate respondents into two groups: those who answered "yes" to at least one part of question 20 and those who answered "no" to all parts. The aim is to identify beforehand a large part of the potential disabled population, in order to focus survey resources on the target group. However, previous surveys have shown that this question will not identify the entire target population. (See Dolson *et al.* 1984 and Dolson *et al.* 1986.)

¹ D. Dolson, K. McClean, J.-P. Morin, and A. Théberge, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6.

The second stage is HALS. Personal interviews are conducted for the "yes" stratum and telephone interviews are conducted for the "no" stratum. From an operational point of view, the interviews are in two parts – the screening questionnaire and the follow-up.

The screening questionnaire is designed to identify respondents for whom the follow-up questionnaire is relevant. The questionnaire for adults covers the seventeen activities of daily living (ADLs) used in the Canadian Health and Disability Survey in 1983 and 1984, repeats Part (a) of question 20 from the Census, and includes a few questions on mental illness and handicaps (see the Appendix). If an affirmative answer is given to at least one of these questions, the interviewer proceeds with the follow-up; if not, the interview is terminated. Part (a) of the Census question is asked again because there may have been a change in status, either because the response in the Census was given by a proxy, or because the respondent has reassessed his or her own condition.

The screening section in the questionnaire for children includes questions on special aids, activity limitations, attendance at a special school and health conditions or problems. A "yes" answer to at least one of these questions prompts a follow-up interview. The Census question is not repeated because all interviews regarding children require a proxy and the question on activity limitations is equivalent to Part (a) of Census question 20.

The second section of this article describes how the population of Canada has been divided into various subpopulations for estimation purposes. The third section covers the HALS sample design. The fourth section deals with the file of geographic information and projected demographic data for 1986 that was used to create the survey frame. The fifth section explains how the sampling was done.

2. POPULATIONS COVERED

Permanent residents of general and psychiatric hospitals, special care centres or institutions for the elderly or chronically ill, institutions for the physically handicapped and orphanages or children's homes are the subject of a distinct part of the survey – namely, HALS (Institutions). This article will look at the part of the survey covering that portion of the Canadian population not covered by HALS (Institutions) and not residing in jails, military camps, young offender facilities, naval vessels, penal or correctional institutions and collective dwellings in the "others" category (for example, circuses and non-religious communes).

Each enumeration area (EA) whose population is not totally excluded from the survey is classified in one of the following five survey frames:

1. Indian reserves where the 1981 Census was conducted using canvassers;
2. Other Indian reserves;
3. Canvasser EAs;
4. EAs in the Whitehorse, Yellowknife, Pine Point, Hay River and Fort Smith SPAs;
5. All other EAs.

The order of priority for belonging to a frame is 1-2-4-3-5. This means that an EA that is an Indian reserve and situated in the Whitehorse SPA is classified as an Indian reserve.

Each EA is divided in two, with the "yes" EA made up of those persons who would answer "yes" to the Census question, and the "no" EA made up of those who would answer "no" to it. A different sample design is used for each of the five survey frames: all of the "yes" EAs and none of the "no" EAs are selected in the first frame; all of the "yes" EAs and a sample of the "no" EAs are selected in the second frame; none of the "no" EAs and a sample of the "yes" EAs are selected in the third frame; all of the EAs are selected in the fourth frame; and a sample of the "yes" EAs and a sample of the "no" EAs are selected in the fifth frame.

3. SURVEY DESIGN

The sampling method presented in this section was used for survey frames three and five. Because our space is limited, the sample design used for the second survey frame will not be described in this article. (For more information on the HALS methodology, see Dolson *et al.* 1986.)

3.1 Sample Design

Each province is divided into subprovincial areas (SPAs), which are themselves divided into enumeration areas (EAs).

Each EA is divided into a “yes” EA and a “no” EA, the first containing those persons who would answer “yes” to Census question 20, the second containing those persons who would answer “no” to that question. In each SPA, the “yes” EAs are stratified into large and small EAs on the basis of the criterion explained in the fourth section of this paper. Persons belonging to a “yes” EA are associated with a stratum and an SPA in addition to their EA, while persons belonging to a “no” EA are associated only with their EA. In each province, the population is subdivided into three age groups: children (under 15 years of age), adults (15 to 64 years of age) and seniors (65 years of age and older).

The sampling method involves using a two-stage stratified sample design for the “yes” EAs in each SPA and a two-stage sample design for the “no” EAs in the province. The primary units are the EAs and the secondary units are the respondents.

All persons who completed Census form 2B in a “yes” EA selected for the sample are interviewed, along with a third of those in the “no” EAs selected.

3.2 Sample Allocation

This sample design must allow us to minimize sampling costs for a given maximum coefficient of variation of the estimates and a given variance for the estimator \hat{B} of the relative bias B . We define B as the ratio of the number of “no” persons with a characteristic of interest in the province, T_0 , to the number of “yes” persons with a characteristic of interest in the province, T_1 . By “no” person, we mean an individual who would answer “no” to all parts of Census question 20, and by “yes” person, an individual who would answer “yes” to at least one part of the question.

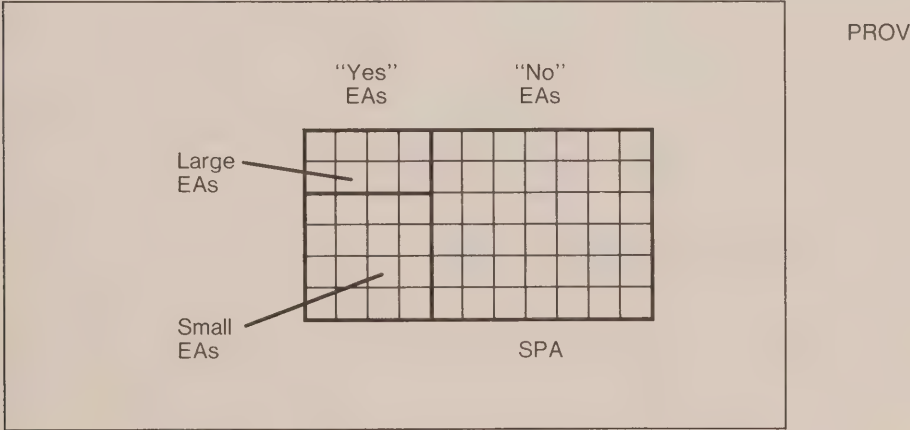


Figure 1. Illustration of Sample Design.

Let N_0 be the number of “no” EAs in the province; N_{jk} , the number of “yes” EAs in stratum j and SPA k in the province; n_0 and n_{jk} , the corresponding sample sizes; and c_0 and c_{jk} , the corresponding unit sampling costs. If we have an N_p SPAs in the province, we therefore want to minimize

$$\sum_{k=1}^{N_p} (c_{1k} n_{1k} + c_{2k} n_{2k}) + c_0 n_0$$

given

$$CV^2(y_k) \leq CV_*^2; \text{Var}(\hat{B}) = \text{Var}_*(\hat{B});$$

$$n_{jk} \leq N_{jk}; n_{2k} = \lambda_k n_{1k}; n_0 \leq N_0$$

$$(j = 1, 2; k = 1, \dots, N_p)$$

where λ_k is the ratio of the expected number of disabled persons in the small EAs to the expected number of disabled persons in the large EAs of SPA k , y_k is the estimated number of “yes” persons who have a characteristic of interest in SPA k , and values marked with an asterisk are constants.

If the sampling fraction in the “yes” EAs is f_1 , M_{ijk} is the number of “yes” persons in EA i of stratum j of SPA k in the province and p_{ijk} is the probability of a characteristic of interest for a “yes” person in EA i of stratum j in SPA k , then

$$E(y_k) = Y_k = \sum_{j=1}^2 \sum_{i=1}^{N_{jk}} M_{ijk} p_{ijk},$$

$$\text{Var}(y_k) = \sum_{j=1}^2 \left\{ \frac{N_{jk}^2}{n_{jk}} \left(1 - \frac{n_{jk}}{N_{jk}} \right) S_{jk}^2 + \frac{N_{jk}}{n_{jk}} \left(\frac{1 - f_1}{f_1} \right) \sum_{i=1}^{N_{jk}} M_{ijk} S_{ijk}^2 \right\},$$

where

$$y_k = \sum_{j=1}^2 \sum_{i=1}^{N_{jk}} \frac{N_{jk}}{n_{jk}} M_{ijk} p_{ijk},$$

$$S_{jk}^2 = \frac{1}{N_{jk} - 1} \sum_{i=1}^{N_{jk}} \left(M_{ijk} p_{ijk} - \left(\sum_{i=1}^{N_{jk}} \frac{M_{ijk} p_{ijk}}{N_{jk}} \right) \right)^2,$$

$$S_{ijk}^2 = \frac{M_{ijk}}{M_{ijk} - 1} p_{ijk} (1 - p_{ijk}).$$

After a few algebraic manipulations, we obtain

$$\begin{aligned} \text{Var}(y_k) = & \frac{1}{n_{1k}} \left\{ N_{1k}^2 S_{1k}^2 + \left(\frac{1 - f_1}{f_1} \right) N_{1k} \sum_{i=1}^{N_{1k}} M_{i1k} S_{i1k}^2 + \frac{N_{2k}^2 S_{2k}^2}{\lambda_k} \right. \\ & \left. + \left(\frac{1 - f_1}{f_1} \right) \frac{N_{2k}}{\lambda_k} \sum_{i=1}^{N_{2k}} M_{i2k} S_{i2k}^2 \right\} - \sum_{j=1}^2 N_{jk} S_{jk}^2. \end{aligned}$$

We can therefore write $CV^2(y_k)$ as

$$CV^2(y_k) = \frac{\text{Var}(y_k)}{Y_k^2} = \frac{A_k}{n_{1k}} - B_k. \quad (3.1)$$

Furthermore, B (the relative bias) and \hat{B} (its estimator) are given by

$$B = \frac{T_0}{T_1} = \frac{\sum_{i=1}^{N_0} M_{i0} p_{i0}}{\sum_{k=1}^{N_p} Y_k},$$

$$\hat{B} = \frac{t_0}{t_1} = \frac{\sum_{i=1}^{n_0} \frac{N_0}{n_0} M_{i0} p_{i0}}{\sum_{k=1}^{N_p} y_k},$$

where M_{i0} is the number of "no" persons in EA i in the province and p_{i0} is the probability of a characteristic of interest for a "no" person in EA i .

Assuming that t_0 and t_1 are independent, then

$$\text{Var}(\hat{B}) = B^2 \left(\frac{\text{Var}(t_0)}{T_0^2} + \frac{\text{Var}(t_1)}{T_1^2} \right). \quad (3.2)$$

After a few algebraic manipulations, if f_0 is the sampling fraction in the "no" EAs, we obtain

$$\text{Var}(t_0) = \frac{1}{n_0} \left\{ N_0^2 S^2 + N_0 \left(\frac{1-f_0}{f_0} \right) \sum_{i=1}^{N_0} M_{i0} S_i^2 \right\} - N_0 S^2$$

where

$$S^2 = \frac{1}{N_0 - 1} \sum_{i=1}^{N_0} \left(M_{i0} p_{i0} - \left(\sum_{i=1}^{N_0} \frac{M_{i0} p_{i0}}{N_0} \right) \right)^2, \quad S_i^2 = \frac{M_{i0}}{M_{i0} - 1} p_{i0} (1 - p_{i0})$$

which can be written in the form

$$\text{Var}(t_0) = \frac{A_0}{n_0} - B_0. \quad (3.3)$$

Furthermore, assuming that the y_k 's are independent, we have

$$\text{Var}(t_1) = \sum_{k=1}^{N_p} \text{Var}(y_k).$$

Using equation (3.1), this expression can be written as

$$\text{Var}(t_1) = \sum_{k=1}^{N_p} \left(\frac{A_k Y_k^2}{n_{1k}} - B_k Y_k^2 \right). \quad (3.4)$$

From (3.2), (3.3) and (3.4), we obtain

$$\begin{aligned} \text{Var}(\hat{B}) &= B^2 \left\{ \left(\frac{A_0}{n_0 T_0^2} - \frac{B_0}{T_0^2} \right) + \sum_{k=1}^{N_p} \left(\frac{A_k Y_k^2}{n_{1k} T_1^2} - \frac{B_k Y_k^2}{T_1^2} \right) \right\} \\ &= \frac{1}{n_0} \left(\frac{B^2 A_0}{T_0^2} \right) + \sum_{k=1}^{N_p} \frac{1}{n_{1k}} \left(\frac{B^2 A_k Y_k^2}{T_1^2} \right) - B^2 \left(\frac{B_0}{T_0^2} + \sum_{k=1}^{N_p} \frac{B_k Y_k^2}{T_1^2} \right) \\ &= \frac{X}{n_0} + \sum_{k=1}^{N_p} \frac{W_k}{n_{1k}} - Z. \end{aligned}$$

The optimization problem can be re-expressed as the problem of minimizing

$$\sum_{k=0}^{N_p} c_k n_k$$

subject to

$$0 < a_k \leq n_k \leq b_k \quad (k = 0, 1, 2, \dots, N_p) \quad (3.5)$$

and

$$\sum_{k=0}^{N_p} d_k / n_k = e \quad (3.6)$$

where, for $k = 1, 2, \dots, N_p$,

$$n_k = n_{1k}, \quad c_k = c_{1k} + c_{2k} \lambda_k, \quad a_k = \frac{A_k}{C V_*^2 + B_k}, \quad b_k = \min(N_{1k}, N_{2k} / \lambda_k).$$

In practice, rather than using $b_k = \min(N_{1k}, N_{2k} / \lambda_k)$ we define

$$b_k = \frac{N_{1k} N_{2k} (1 + \lambda_k)}{\lambda_k^2 N_{1k} + N_{2k}},$$

then, if $n_k > N_{1k}$, sample sizes are given by $n_{1k} = N_{1k}$ and

$$n_{2k} = \lambda_k n_k + \frac{(n_k - N_{1k}) N_{2k}}{N_{1k} \lambda_k},$$

while, if $\lambda_k n_k > N_{2k}$ sample sizes are given by $n_{2k} = N_{2k}$ and

$$n_{1k} = n_k + \frac{(\lambda_k n_k - N_{2k}) N_{1k} \lambda_k}{N_{2k}}.$$

Thus, we consider $N_{2k} / (N_{1k} \lambda_k)$ small EAs to be equivalent to one large EA. On average, there are as many disabled persons in one large EA as in $N_{2k} / (N_{1k} \lambda_k)$ small EAs.

Proceeding in this way, it is not always true that $n_{2k} = \lambda_k n_{1k}$. However, we avoid CVs higher than target values, when, for example, small EAs remain to be observed (even if all the large EAs have been selected).

For some values of k , it is possible that $a_k \geq b_k$. If this is the case, we set $n_k = b_k$. Let

$$\begin{aligned} E_1 &= \{k = 0, 1, 2, \dots, N_p \mid n_k = a_k\}, \\ E_2 &= \{k = 0, 1, 2, \dots, N_p \mid n_k = b_k > a_k\}, \\ E_3 &= \{k = 0, 1, 2, \dots, N_p \mid a_k < n_k < b_k\}, \\ E_4 &= \{k = 0, 1, 2, \dots, N_p \mid n_k = b_k \leq a_k\}. \end{aligned}$$

The solution exists if

$$\sum_{k=0}^{N_p} d_k / b_k \leq e,$$

and it takes the form

$$n_k = \begin{cases} a_k & (k \in E_1) \\ b_k & (k \in E_2 \cup E_4) \\ K (d_k / c_k)^{1/2} & (k \in E_3) \end{cases} \tag{3.7}$$

where

$$K = \frac{\sum_{k \in E_3} (d_k / c_k)^{1/2}}{e - \sum_{k \in E_1} d_k / a_k - \sum_{k \in E_2 \cup E_4} d_k / b_k}, \tag{3.8}$$

since the n_k ($k \in E_3$) minimize $\sum_{k \in E_3} c_k n_k$, subject to the constraint

$$\sum_{k \in E_3} d_k / n_k = e - \sum_{k \in E_1} d_k / a_k - \sum_{k \in E_2 \cup E_4} d_k / b_k.$$

What are the sets E_1 , E_2 , E_3 and E_4 corresponding to the solution? Set E_4 is easy to determine. We must have

$$a_k < (d_k / c_k)^{1/2} K < b_k \quad (k \in E_3), \quad (d_k / c_k)^{1/2} K \geq b_k \quad (k \in E_2),$$

$$(d_k / c_k)^{1/2} K \leq a_k \quad (k \in E_1). \quad (3.9)$$

Determining the sets involves trying each of the possibilities for E_1 , E_2 and E_3 until a value for k which satisfies (3.9) is obtained. To reduce the number of possibilities to be examined, note that, if for $k' \geq k$,

$$b'_k (c'_k / d'_k)^{1/2} \geq b_k (c_k / d_k)^{1/2} \quad (k, k' \in \{0, 1, \dots, N_p\}), \quad (3.10)$$

then there is a k^* such that $E_2 = \{0, 1, 2, \dots, k^*\}$, or $E_2 = \{ \}$, while, if for $k' \geq k$,

$$a'_k (c'_k / d'_k)^{1/2} \geq a_k (c_k / d_k)^{1/2} \quad (k, k' \in \{0, 1, \dots, N_p\}), \quad (3.11)$$

then there is a k^{**} such that $E_1 = \{k^{**}, k^{**} + 1, \dots, N_p\}$ or $E_1 = \{ \}$.

3.3 Parameter Estimation

To calculate the optimum sample allocation, the following quantities must be determined:

P_1 = proportion of HALS screened-in individuals who replied "yes" to Census question 20,

P_2 = proportion of HALS screened-out individuals who replied "yes" to Census question 20, and

P_3 = proportion of HALS screened-in individuals who replied "no" to Census question 20.

Since these parameters cannot be computed directly using data from the Canadian Health and Disability Survey, a test called the "calibration study" was carried out in September and October 1985.

Census question 20 was included, without abbreviation, as a supplementary question in the September Labour Force Survey (LFS). It was asked to a sample of approximately 36,000 individuals. The questions on the 17 ADLs and a question on mental handicaps were added as a supplement to the October LFS and were asked of the same individuals.

For each five-year age group, the weighted values from the calibration study were used to estimate the probability of an affirmative response, $P(\text{yes})$, to Census question 20. The HALS screening questionnaire differs from that used in the calibration study. In HALS, there are more questions on mental and psychological problems and part (a) of Census question 20 is asked again. Therefore, we did not depend on the calibration study alone to calculate the parameters.

4. 1986 GEOGRAPHIC AND DEMOGRAPHIC FILE

4.1 Description of Available Information

When the sample allocation was done in the spring of 1986, the following information was available for use in calculation of population projections by age group and EA:

- 1. population projections by age group and province in 1986;
- 2. estimated population by age group and CD in 1984;
- 3. population by age group and EA in 1981;
- 4. conversion file to establish the correspondence between the 1981 and 1986 EAs;
- 5. estimated numbers of dwellings by EA in 1986.

The conversion file is structured according to the concept of equivalent sets. Each equivalent set is the smallest region consisting of EAs that has not had its boundaries altered. For example, if three 1981 EAs were reorganized as two 1986 EAs, the group of three 1981 EAs (or the group of two 1986 EAs) is an equivalent set.

The four methods described in the next subsection are designed to produce population projections by age group and by equivalent set in 1986. If an equivalent set is made up of several 1986 EAs, the projected population for the equivalent set can be divided proportionally among the EAs using the estimated numbers of dwellings by EA in 1986.

4.2 Estimation Methods

For province p , let

- $ES_{l,k}$ = the l -th equivalent set of the k -th CD ($l = 1, 2, \dots, N_k; k = 1, 2, \dots, N_p$),
- $ES_{l,k;81}(j)$ = population of $ES_{l,k}$ in the j -th age group in 1981 ($j = 1, 2, \dots, 16$),
- $CD_{k;84}(j)$ = estimated population of the k -th CD in the j -th age group in 1984,
- $\hat{P}_{86}(j)$ = projected population in the j -th age group in the province in 1986.

For the three methods that follow, the first step is to calculate $\hat{CD}_{k;86}(j)$, the projected population of the j -th age group in the k -th CD in 1986. We assume there exists K'_j ($j = 1, 2, \dots, 16$) such that

$$\hat{CD}_{k;86}(j) = K'_j(\hat{CD}_{k;84}(j)) \quad (k = 1, 2, \dots, N_p; j = 1, 2, \dots, 16),$$

$$\sum_{k=1}^{N_p} \hat{CD}_{k;86}(j) = \hat{P}_{86}(j) \quad (j = 1, 2, \dots, 16).$$

This implies that

$$\hat{CD}_{k;86} = \frac{\hat{P}_{86}(j) CD_{k;84}(j)}{\sum_{k=1}^{N_p} CD_{k;84}(j)}.$$

The first method of estimating $ES_{l,k;86}(j)$ involves assuming the existence of K_j ($j = 1, \dots, 16$) such that

$$\hat{ES}_{l,k;86}(j) = K_j ES_{l,k;81}(j) \quad (l = 1, \dots, N_k; j = 1, \dots, 16),$$

$$\sum_{l=1}^{N_k} \hat{ES}_{l,k;86}(j) = \hat{CD}_{k;86}(j) \quad (j = 1, 2, \dots, 16).$$

We will say that this method uses the simple model. We obtain

$$\hat{ES}_{l,k;86}(j) = \frac{\hat{CD}_{k;86}(j) ES_{l,k;81}(j)}{\sum_{l=1}^{N_k} ES_{l,k;81}(j)} \quad (l = 1, \dots, N_k; j = 1, \dots, 16).$$

With this simple model, the estimated total population of $ES_{l,k}$ in 1986 is

$$\sum_{j=1}^{16} \frac{\hat{CD}_{k;86}(j) ES_{l,k;81}(j)}{\sum_{l=1}^{N_k} ES_{l,k;81}(j)}.$$

If one thinks that a better estimate, $\hat{ES}_{l,k;86}(tot)$ of this quantity can be produced by independent means (for example, using the estimated number of dwellings in $ES_{l,k}$ in 1986), then more elaborate models can be used to estimate $ES_{l,k;86}(j)$. The multiplicative model is specified by the following equations:

$$\hat{ES}_{l,k;86}(j) = K_j (ES_{l,k;81}(j)) + e'_l \quad (l = 1, \dots, N_k; j = 1, \dots, 16),$$

$$\sum_{l=1}^{N_k} \hat{ES}_{l,k;86}(j) = K (\hat{CD}_{k;86}(j)) \quad (j = 1, \dots, 16),$$

$$\sum_{l=1}^{N_k} e'_l = 0,$$

$$\sum_{j=1}^{16} \hat{ES}_{l,k;86}(j) = \hat{ES}_{l,k;86}(tot) \quad (l = 1, \dots, N_k).$$

One can interpret e_l as the net intra-CD migration for the l -th equivalent set.

The third model, called the additive model, is given by the following equations:

$$\hat{ES}_{l,k;86}(j) = ES_{l,k;81}(j) + e_l + f_j \quad (l = 1, \dots, N_k; j = 1, \dots, 16),$$

$$\sum_{l=1}^{N_k} \hat{ES}_{l,k;86}(j) = \hat{CD}_{k;86}(j) + D \quad (j = 1, \dots, 16),$$

$$\sum_{l=1}^{N_k} e_l = D,$$

$$\sum_{j=1}^{16} \hat{ES}_{l,k;86}(j) = \hat{ES}_{l,k;86}(tot) \quad (l = 1, \dots, N_k).$$

This model involves the assumption that the population increases (or decreases) for each age group in each of the equivalent sets in a CD can be decomposed into two terms – one which depends only on the equivalent set and not on age (e_l), and one which depends only on age and not on the equivalent set (f_j).

A final trivial model involves simply formulating

$$\hat{ES}_{l,k;86}(j) = ES_{l,k;81}(j) \quad (l = 1, 2, \dots, N_k; j = 1, \dots, 16).$$

4.3 Evaluation of Estimation Methods

The four methods were evaluated using data for the period 1976-1981. We used the 1976 projection of the population by age group and province in 1981, ($\hat{P}_{81}(j)$), the population by age group and EA in 1976, a 1976-1981 conversion file and the pre-Census estimate of the number of dwellings per EA in 1981. Since there are no estimates for population by age group and CD in 1979 (the equivalent of $CD_{k;84}(j)$), we set

$$\hat{CD}_{k;81} = \frac{\hat{P}_{81}(j) \hat{CD}_{k;84}(j)}{\sum_{k=1}^{N_p} \hat{CD}_{k;84}(j)}.$$

For $\hat{ES}_{l,k;81}(tot)$, which is needed for the multiplicative and additive models, we used

$$\hat{ES}_{l,k;81}(tot) = \frac{\sum_{j=1}^{16} ES_{l,k;76}(j) \sum_{j=1}^{16} \hat{CD}_{k;81}(j)}{\sum_{l=1}^{N_k} \sum_{j=1}^{16} ES_{l,k;76}(j)}.$$

Table 1
Comparison of the Four Methods

Prov.	EFF_S	EFF_M	EFF_A
Nfld.	0.890	0.891	0.887
P.E.I.	0.903	0.914	0.919
N.S.	0.960	0.972	0.912
N.B.	0.870	0.868	0.884
Que.	0.778	0.764	0.818
Ont.	0.932	0.930	0.916
Man.	0.892	0.904	0.912
Sask.	0.732	0.749	0.801
Alta.	0.818	0.827	0.860
B.C.	0.713	0.716	0.775
Yukon	0.770	0.768	0.840
N.W.T.	1.252	1.246	1.157

For each province p , an efficiency measure was calculated for the simple, multiplicative and additive models relative to the trivial model:

$$EFF_m = \frac{\sum_{k=1}^{N_p} \sum_{l=1}^{N_k} \sum_{j=1}^{16} \left((\hat{ES}_{l,k;81}^{(m)}(j) - ES_{l,k;81}(j))^2 \right)}{\sum_{k=1}^{N_p} \sum_{l=1}^{N_k} \sum_{j=1}^{16} \left((\hat{ES}_{l,k;81}^{(T)}(j) - ES_{l,k;81}(j))^2 \right)} \quad (m = S, M, A),$$

where $\hat{ES}_{l,k;81}^{(m)}(j)$ with $m = S, M, A$ and T are the projections obtained by means of the simple, multiplicative, additive and trivial models respectively. Some values obtained are given in Table 1.

The simple model gives the worst results for one province and one territory, the multiplicative model for two provinces and the additive model for seven provinces and one territory.

The simple model is the best for five provinces, while the multiplicative model is best for two provinces and one territory and the additive model is best for three provinces and one territory.

Since the simple model also has, as its name implies, the advantage of simplicity, it is the one that was chosen.

4.4 Method of Stratification by Enumeration Area Size

If simple random sampling were used to select EAs within each subprovincial area (SPA), disabled persons belonging to an EA with many disabled residents would have less chance of being selected than those in a small EA – that is, an EA with few disabled persons. To avoid excessive differences in selection probabilities, the population of EAs in each SPA is stratified according to the number of disabled persons in the EAs, and then proportional allocation is used. With proportional allocation, the number of EAs selected is proportional to the number of disabled persons for each stratum.

Using the results of earlier surveys, a link was established between the age distribution of the population of an EA and the number of disabled persons expected in the EA. Since the number of disabled persons is unknown, the variable used for stratification and sample allocation is the expected number of disabled persons.

In the case under consideration here, there are only two strata – one for large EAs and one for small EAs. Since proportional allocation is being used, we employed a criterion found in Raj (1968) to determine the optimum dividing line between large and small EAs. This criterion gives the optimum dividing line as the average of the average size of the small EAs and the average size of the large EAs.

5. SAMPLE SELECTION

It was necessary to draw samples for the three populations (children, adults and seniors) among the large and small “yes” EAs of each SPA, both for frame three and for frame five, and among the “no” EAs of each province for frame five. When an SPA contained fewer than two large EAs or fewer than two small EAs, we selected all of the EAs in that SPA for the three populations. The “yes” and “no” samples were created independently, using the one-pass algorithm described by Bebbington (1975). The samples from the three populations for the “yes” and “no” components were nested to minimize the total number of EAs selected.

The following table shows the sizes obtained for the samples by province for each age group.

Table 2
Sample Sizes by Province and Age Group

Province	Children		Adults		Seniors	
	Number of “yes” EAs selected	Number of “no” EAs selected	Number of “yes” EAs selected	Number of “no” EAs selected	Number of “yes” EAs selected	Number of “no” EAs selected
Nfld.	880	136	405	154	476	173
P.E.I.	242	242	111	217	82	166
N.S.	1257	157	434	130	438	115
N.B.	1142	162	459	146	453	138
Que.	4749	153	1070	114	1488	133
Ont.	6085	158	1304	116	1542	120
Man.	1082	203	457	169	367	144
Sask.	2291	265	942	241	921	193
Alta.	2762	190	909	176	1389	222
B.C.	3117	170	752	125	948	119

6. DISCUSSION

The postcensal survey is a relatively new survey method that will no doubt undergo extensive development in the next few years. This type of survey allows for a great deal of flexibility in data collection and use of large samples scattered throughout the country, with reasonable costs and timeliness. The Health and Activity Limitation Survey is the first postcensal survey of its size in Canada.

The sample design presented in this article is an attempt to maximize use of the opportunities offered by the postcensal approach, with optimum use of the available resources. One of the major problems inherent in the proposed method is control of sample size. Sample allocation is determined before the census is taken; this means that all calculations must be done using projections based on the previous census. In this context, the actual size of a sample made up of a group of small areas selected on the basis of the projection results may vary considerably from its expected size.

Therefore, on the one hand, one may obtain a sample that is inadequate with respect to the quality requirements for the estimates. On the other hand, the resources allocated to data collection may be exceeded. In order to prevent these problems, we implemented the following strategy. A target number of interviews for each population was calculated for the "yes" sample. This number was based on the sample size required to produce estimates that would satisfy our quality criteria. However, for the reasons mentioned above, we selected more EAs than were necessary to obtain the target number of interviews. For reasons of cost, if the real number of interviews to be conducted, as calculated in the field, was higher than the target number, a sub-sample of EAs were excluded from the survey. Only for the Halifax Regional Office (covering Prince Edward Island, Nova Scotia and New Brunswick) was the number of interviews in the "yes" sample substantially higher than the target number. The decision was therefore made to exclude certain EAs from this part of the sample. In order to know which EAs would be excluded, it was necessary to know the target number and the real number of interviews for each EA. For 40 per cent of the EAs, the real number of interviews had to be imputed since this information was not available in time.

For this imputation, the total real number of interviews was known for each census commissioner district. The portion of this total not already allocated to EAs with known numbers of interviews was distributed among the EAs requiring imputation, in proportion to the target number of interviews.

We then calculated, for each population, the difference between the real number and the target number of interviews for each of the two strata of each SPA. A positive difference (real-target) indicated a population for which some EAs could be excluded from the survey. In each stratum, the EAs were divided into three groups (1, 2 and 3), in accordance with whether they had been selected for three, two or only one of the populations respectively. The EA file was then sorted by stratum and by group in ascending order, with the order of the EAs within each group being random. Each EA was considered successively and was suppressed for the three populations if:

- 1) a positive difference remained non-negative after suppression of the EA;
- 2) a negative difference was not further reduced.

In this way, each positive difference was reduced to a number as close as possible to zero, considering the random order of the EAs.

ACKNOWLEDGEMENTS

The authors would like to thank D.A. Binder for his contribution with regard to the sample design and P. Reed for the computer work involved in the sample allocation. The authors would also like to thank the referees for their useful comments.

APPENDIX**Question 20 of Census Form 2B**

20. a) Are you limited in the kind or amount of activity that you can do because of a long-term physical condition, mental condition or health problem: (See Guide)

At home?

☐ No, I am not limited

☐ Yes, I am limited

At school or at work?

☐ No, I am not limited

☐ Yes, I am limited

☐ Not applicable

In other activities, e.g., transportation to or from work, leisure time activities?

☐ No, I am not limited

☐ Yes, I am limited

b) Do you have any long-term disabilities or handicaps?

☐ No

☐ Yes

Screening Questions for HALS (Questionnaire for Adults)

1. Do you have any trouble hearing what is said in a normal conversation with one other person?
2. Do you have any trouble hearing what is said in a group conversation with at least three other people?
4. Do you have any trouble reading ordinary newsprint, with glasses if normally worn?
5. Do you have any trouble seeing clearly the face of someone from 12 feet/4 metres (example: across a room), with glasses if normally worn?
7. Do you have any trouble speaking and being understood?
8. Do you have any trouble walking 400 yards/400 metres without resting (about three city blocks)?
9. Do you have any trouble walking up and down a flight of stairs (about 12 steps)?
10. Do you have any trouble carrying an object of 10 pounds for 30 feet/5 kg for 10 metres (example: carrying a bag of groceries)?
11. Do you have any trouble moving from one room to another?
12. Do you have any trouble standing for long periods of time, that is, more than 20 minutes? Remember, I am asking about problems expected to last 6 months or more.
13. When standing do you have any trouble bending down and picking up an object from the floor (example: a shoe)?
14. Do you have any trouble dressing and undressing yourself?
15. Do you have any trouble getting in and out of bed?
16. Do you have any trouble cutting your own toenails?

17. Do you have any trouble using your fingers to grasp or handle?
18. Do you have any trouble reaching in any direction (example: above your head)?
19. Do you have any trouble cutting your own food?
20. Because of a long-term physical condition or health problem, that is, one that is expected to last 6 months or more, are you limited in the kind or amount of activity you can do . . .
(i) at home? (ii) at school or at work? (iii) in other activities such as travel, sports, or leisure?
21. Has a school or health professional ever told you that you have a learning disability?
22. From time to time, everyone has trouble remembering the name of a familiar person, or learning something new, or they experience moments of confusion. However, do you have any ongoing problems with your ability to remember or learn?
23. Because of a long-term emotional, psychological, nervous, or mental health condition or problem, are you limited in the kind or amount of activity you can do?
(i) at home? (ii) at school or at work? (iii) in other activities such as travel, sports, or leisure?

REFERENCES

- BEBBINGTON, A.C. (1975). A simple method of drawing a sample without replacement. *Applied Statistics*, 24, 136.
- CARTER, R.G., GILES, P.D., and SHERIDAN, M.J. (1982). Description and rationale for the screen tests for the January 1983 Disability Survey. Disability Data Development Project, Health Division, Statistics Canada.
- HOUSE OF COMMONS (1981). Obstacles, Report of the special committee on the disabled and handicapped. Ottawa.
- COCHRAN, W.G. (1977). *Sampling Techniques*, (3rd ed.). New York: John Wiley.
- DOLSON, D., GILES, P., and MORIN, J.-P. (1984). A Methodology for surveying disabled persons using a supplement to the Labour Force Survey. *Survey Methodology*, 10, 187-197.
- DOLSON, D., McCLEAN, K., MORIN, J.-P., and THÉBERGE, A. (1986). Methodology report of HALS. Working Paper, Statistics Canada.
- GRABOWIECKI, F. (1982). Discussion of the target population for the Disability Survey. Disability Data Development Project, Health Division, Statistics Canada.
- GRABOWIECKI, F. (1983). Content of Statistics Canada's Disability Survey. Technical Report, Health Division, Statistics Canada.
- LAZARUS, G., and NESICH, R. (1985). A report on the methodology of the Canadian Health and Disability Survey. Working Paper, Statistics Canada.
- McDOWELL, I. (1981). An examination of the OECD survey questions in a Canadian Study. *Revue d'épidémiologie et de santé publique*, 29, 412-429.
- MORIN, J.-P., and DOWLER, L. (1986). Proposition d'une méthodologie pour l'ESLA-institutions. Working Paper, Statistics Canada.
- MORIN, J.-P. (1986). Comparaison initiale de l'ESIC et de l'ESG. Working Paper, Statistics Canada.
- WORLD HEALTH ORGANIZATION (1980). International classification of impairments, disabilities and handicaps. Geneva, Switzerland.
- RAJ, D. (1968). *Sampling Theory*. New York: McGraw-Hill.
- WILSON, R.W., and McNEIL, J.M. (1981). Preliminary analysis of OECD disability on the pretest of the Post-censal Disability Survey. *Revue d'épidémiologie et de santé publique*, 29, 469-475.

Comparison of Estimators of Population Total in Two-Stage Successive Sampling Using Auxiliary Information

F.C. OKAFOR¹

ABSTRACT

Singh and Srivastava (1973) proposed a linear unbiased estimator of the population mean when sampling on successive occasions using several auxiliary variables whose known population means remain unchanged for all occasions. In this paper, three composite estimators T_1 , T_2 and T_3 , each utilising an auxiliary variable whose known population mean changes from one occasion to the next, are presented for the estimation of the current population total. The proposed estimators are compared with the ordinary estimator, T_0 , and the usual successive sampling estimator, T' , of the current population total without the use of auxiliary information. We find that using auxiliary information in conjunction with successive sampling does not always uniformly produce a gain in efficiency over T_0 or T' . However, when applied to a survey of teak plantations to estimate the mean height of teak trees, T_1 , T_2 and T_3 proved more efficient than T_0 and T' .

KEY WORDS: Successive occasion; Partial matching; Auxiliary variate.

1. INTRODUCTION

The theory and practice of surveying the same population at different points in time – technically called repetitive sampling or sampling over successive occasions – have been given considerable attention by some survey statisticians. The main objective of sampling on successive occasions is to estimate some population parameters (total, mean, ratio, etc) for the most recent occasion as well as changes in these parameters from one occasion to the next.

The theory of successive sampling was initiated by Jessen (1942). Many authors have since contributed, especially in the estimation of population means. Among them are Singh (1968), Abraham et al (1969), Kathuria and Singh (1971), and Kathuria (1975), to mention but a few.

Singh (1968) was the first to extend the theory of unistage sampling to two-stage sampling on successive occasions. He considered the sampling scheme in which, on the second occasion, a fraction λ of the first stage units (FSUs) selected on the previous occasion is retained, along with their selected second stage units (SSUs), and a fraction μ ($\lambda + \mu = 1$) selected afresh. He then obtained a minimum variance unbiased estimator of the population mean on the current occasion.

Abraham et al (1969) considered the situation in which partial matching of units was carried out at both stages. Units were selected by simple random sampling without replacement (SRSWOR). Kathuria (1975) modified this by using probability proportional to size and with replacement (PPSWR) for selection of the FSUs, and proposed a linear composite estimator for the population mean on the current occasion.

¹ F.C. Okafor, Department of Statistics, University of Ibadan, Ibadan, Nigeria.

When an auxiliary variable is highly correlated with the characteristic under study, the estimate of the population mean (total) of this characteristic can be improved using the auxiliary variable. Singh and Srivastava (1973) used auxiliary information to improve on the estimator of Singh (1968). They obtained a linear unbiased estimator of the population mean on the most recent occasion using several auxiliary variables whose population means are known and are the same for all occasions. Kathuria (1978) developed this study further by assuming that the population mean of the auxiliary variate is not known. He used a double sampling technique to estimate first the population mean of the auxiliary variate and then the mean of the characteristic under study.

In their contributions, Singh and Srivastava (1973) and Kathuria (1978) assumed that the necessary information on the auxiliary variables can be obtained from the respondents or reporting units (SSUs). This is not generally the case. It may happen that the information on the auxiliary variable is too distorted to be useful because of the sensitive nature of the question, or the respondents may refuse outright to supply any information. Alternatively, the information on the auxiliary variate may not be collected because the required question is not included in the questionnaire.

Singh and Srivastava also assumed that the known population total of the auxiliary variable is the same for all occasions. This may not be true in practice. If the population total of the main characteristic changes from one occasion to the next, there is every likelihood that the population total of any other variable correlated with it will also vary.

In this paper three composite estimators of the population total using auxiliary information and a two-stage successive sampling scheme are proposed. The performances of the three estimators are compared empirically and they are also applied to a survey of teak plantations to estimate the mean height of teak trees.

2. SAMPLING FOR TWO OCCASIONS

For all three proposed estimators, we assume that the population total of the auxiliary variable changes on the second occasion.

The estimators of the population total (mean) based on the partial matching scheme are better than the ordinary estimators of the population total (mean) without partial matching. Therefore, it is expected that the proposed estimators T_1 , T_2 and T_3 will perform better than the ordinary population total estimator, T_o , and the estimator based on the partial matching scheme without the use of auxiliary information, T' .

In deriving these estimators, we assume that:

- (i) the sample size is constant on each occasion;
- (ii) the normed size measure P_i for the i^{th} first stage unit (FSU) is fixed for each occasion;
- (iii) N and M_i , population sizes for the FSUs and the second stage units (SSUs) within the i^{th} FSU respectively, are constant for the two occasions;
- (iv) the population total (mean) of the auxiliary variate is known.

Assumptions (i) – (iii) apply to T' , T_1 , T_2 and T_3 ; (iv) applies to T_1 , T_2 and T_3 , but not to T' and T_o .

On the first occasion, a sample S_1 of n FSUs is selected with probability proportional to size and with replacement (PPSWR) using P_i as normed size measure for the i^{th} ($i = 1, 2, \dots, N$) unit. For the selection of SSUs, we adopt the method due to Cochran

(1977, p. 306), which stipulates that if the i^{th} FSU in S_1 is drawn θ_i times ($i = 1, 2, \dots, n$), we select θ_i independent subsamples of size m_i from the M_i SSUs.

On the second occasion, we select a sample of λn ($0 < \lambda < 1$) FSUs from S_1 by simple random sampling without replacement (SRSWOR). The SSUs selected on the first occasion are retained for each of these λn matched FSUs. Then, a fresh sample of μn ($\mu = 1 - \lambda$) FSUs is selected independently from the N FSUs by PPSWR, with P_i as normed size measure for the i^{th} FSU. In each of the μn FSUs, the SSUs are selected as on the first occasion.

3. NOTATION

We define $y_{ij}(x_{ij})$ as the value of the study variate for the j^{th} SSU in the i^{th} FSU on the current (previous) occasion. In addition, z_{hij} is defined as the value of the auxiliary variate for the j^{th} SSU in the i^{th} FSU on the h^{th} occasion ($h = 1, 2$). The sample means for SSUs in the i^{th} FSU are

$$\bar{x}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} x_{ij}, \bar{y}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij} \text{ and } \bar{z}_{hi} = \frac{1}{m_i} \sum_{j=1}^{m_i} z_{hij}.$$

The population total for the i^{th} FSU and the overall population total for the auxiliary variate are

$$Z_{hi} = \sum_{j=1}^{M_i} z_{hij} \text{ and } Z_h = \sum_{i=1}^N Z_{hi}.$$

We define additional notation as follows:

$$S_b^2(y) = \sum_{i=1}^N P_i \left(\frac{Y_i}{P_i} - Y \right)^2 \text{ is the between - FSU variance;}$$
$$S_w^2(y) = \sum_{i=1}^N \frac{M_i^2}{P_i} \left(\frac{1}{m_i} - \frac{1}{M_i} \right) S_{wi}^2(y) \text{ is the variance among SSUs within the FSUs;}$$
$$S_{wi}^2(y) = \frac{1}{M_i - 1} \sum_{j=1}^{M_i} (y_{ij} - \bar{y}_i)^2 \text{ is the variance among the SSUs in the } i^{th} \text{ FSU;}$$
$$S^2(y) = S_b^2(y) + S_w^2(y);$$

$$C_b(x,y) = \rho_b S_b(x) S_b(y) \text{ is the between-FSU covariance of } x \text{ and } y;$$
$$C_w(x,y) = \rho_w S_w(x) S_w(y) \text{ is the covariance of } x \text{ and } y \text{ among SSUs within the FSUs;}$$
$$C(x,y) = C_b(x,y) + C_w(x,y).$$

The between- and within-FSU correlation coefficients between x and y are respectively ρ_b and ρ_w .

4. ESTIMATORS FOR THE POPULATION TOTAL AND THEIR OPTIMUM VARIANCES

4.1 Case (i)

The first estimator of the population total, Y , on the second occasion is used when information on the auxiliary variable is not available but the FSU population total of the auxiliary variable is available for the selected FSUs. It is given as

$$T_1 = \theta(1) T_m(1) + (1 - \theta(1)) T_u(1) \quad (4.1)$$

$\theta(1)$ is a constant chosen so that the variance of T_1 , $V(T_1)$, attains a minimum; while

$$\begin{aligned} T_m(1) = & \frac{1}{\lambda n} \sum_{i=1}^{\lambda n} \left\{ \frac{M_i \bar{y}_i}{P_i} - k(1) \left(\frac{Z_{2i}}{P_i} - Z_2 \right) \right\} \\ & - b(1) \left[\frac{1}{\lambda n} \sum_{i=1}^{\lambda n} \left\{ \frac{M_i \bar{x}_i}{P_i} - k(1) \left(\frac{Z_{1i}}{P_i} - Z_1 \right) \right\} \right. \\ & \left. - \frac{1}{n} \sum_{i=1}^n \left\{ \frac{M_i \bar{x}_i}{P_i} - k(1) \left(\frac{Z_{1i}}{P_i} - Z_1 \right) \right\} \right] \end{aligned}$$

is the difference estimator of Y based on the matched sample;

$$T_u(1) = \frac{1}{n\mu} \sum_{i=1}^{n\mu} \left\{ \frac{M_i \bar{y}_i}{P_i} - k(1) \left(\frac{Z_{2i}}{P_i} - Z_2 \right) \right\}$$

is the estimator for Y based on the unmatched sample; and $k(1)$ and $b(1)$ are known constants.

For this estimator, it is assumed that the population total of the auxiliary variate, Z_i , is available for each selected FSU on each occasion. The overall population total, Z , is also available on each occasion. No additional information on the auxiliary variate is obtained from the respondents (SSUs).

Now by minimizing $V(T_1)$ with respect to $\theta(1)$ and solving, the optimum value of $\theta(1)$ becomes

$$\theta_0(1) = \lambda A_2(1) / \Delta(1)$$

where

$$A_2(1) = S^2(y) + k^2(1) S_b^2(z_2) - 2k(1) C_b(z_2, y),$$

$$\Delta(1) = A_2(1) + \mu^2 \{ b^2(1) A_1(1) - 2b(1) \beta(1) \}.$$

The optimum value of $k(1)$ is obtained by minimizing $V(T_u(1))$ with respect to $k(1)$. This gives $k_0(1) = C_b(z_2, y) / S_b^2(z_2)$.

It can be shown that the optimum $V(T_1)$ for a given λ , following the method adopted by Jessen (1942), is

$$V_0(T_1) = \frac{1}{n} [A_2(1) + \mu \{b^2(1)A_1(1) - 2b(1)\beta(1)\}] A_2(1) / \Delta(1) \quad (4.2)$$

where

$$A_1(1) = S^2(x) + k^2(1) S_b^2(z_1) - 2k(1) C_b(z_1, x),$$

$$\beta(1) = C(x, y) + k^2(1) C_b(z_1, z_2) - k(1) \{C_b(x, z_2) + C_b(z_1, y)\},$$

$$\Delta(1) = A_2(1) + \mu^2 \{b^2(1) A_1(1) - 2b(1) \beta(1)\}.$$

Minimizing the variance of $T_m(1)$, the optimum $b(1)$ is

$$b_0(1) = \beta(1) / A_1(1).$$

If $b_0(1)$ is substituted in (4.2), the optimum variance becomes

$$V_0(T_1) = \frac{1}{n} \left[\frac{A_1(1) A_2(1) - \mu \beta^2(1)}{A_1(1) A_2(1) - \mu^2 \beta^2(1)} \right] A_2(1). \quad (4.3)$$

By minimizing $V_0(T_1)$ in (4.2) with respect to μ , the optimum matching fraction boils down to $\lambda_0 = 1 - \mu_0$ where

$$\mu_0 = A_2(1) [A_2(1) + \{A_2^2(1) + A_2(1) (b^2(1)A_1(1) - 2b(1)\beta(1))\}^{1/2}]^{-1}. \quad (4.4)$$

If $A_2(1) = A_1(1)$, i.e. the population variability is the same on both occasions, the expression in (4.3) yields

$$V_0(T_1) = \frac{1}{n} \left[\frac{A^2(1) - \mu \beta^2(1)}{A^2(1) - \mu^2 \beta^2(1)} \right] A(1) \quad (4.5)$$

while the optimum matching fraction, μ_0 (given in (4.4)), with $b_0(1)$ substituted for $b(1)$ becomes

$$\mu_0 = A(1) [A(1) + \{A^2(1) - \beta^2(1)\}^{1/2}]^{-1}. \quad (4.6)$$

When μ_0 is substituted in (4.5) the variance works out as

$$V_0(T_1) = \frac{1}{2n} [A(1) + \{A^2(1) - \beta^2(1)\}^{1/2}]. \quad (4.7)$$

4.2 Case (ii)

The second estimator is the usual one in which information is obtained on both the main and auxiliary characteristic from the reporting units and the population total of the auxiliary characteristic is known.

It is written as

$$T_2 = \theta(2) T_m(2) + (1 - \theta(2)) T_u(2), \quad (4.8)$$

where

$$\begin{aligned} T_m(2) = & \frac{1}{\lambda n} \sum_{i=1}^{\lambda n} \left\{ \frac{M_i \bar{y}_i}{P_i} - k(2) \left(\frac{M_i \bar{z}_{2i}}{P_i} - Z_2 \right) \right. \\ & - b(2) \left[\frac{1}{\lambda n} \sum_{i=1}^{\lambda n} \left\{ \frac{M_i \bar{x}_i}{P_i} - k(2) \left(\frac{M_i \bar{z}_{1i}}{P_i} - Z_1 \right) \right\} \right. \\ & \left. \left. - \frac{1}{n} \sum_{i=1}^n \left\{ \frac{M_i \bar{x}_i}{P_i} - k(2) \left(\frac{M_i \bar{z}_{1i}}{P_i} - Z_1 \right) \right\} \right] \right\}, \end{aligned}$$

and

$$T_u(2) = \frac{1}{n\mu} \sum_{i=1}^{n\mu} \left\{ \frac{M_i \bar{y}_i}{P_i} - k(2) \left(\frac{M_i \bar{z}_{2i}}{P_i} - Z_2 \right) \right\}.$$

Here the overall population total of the auxiliary variate is known on both occasions. In addition, information on the auxiliary variate, z_{ij} , is obtained for every SSU in the sample. This is the usual way of using the auxiliary information in a two-stage design described in the literature. It can be shown that the optimum variance of T_2 is

$$V_0(T_2) = \frac{1}{n} [A_2(2) + \mu \{b^2(2)A_1(2) - 2b(2)\beta(2)\}] A_2(2) / \Delta(2) \quad (4.9)$$

and the optimum weight is

$$\theta_0(2) = \lambda A_2(2) / \Delta(2)$$

where

$$A_2(2) = S^2(y) + k^2(2) S^2(z_2) - 2k(2) C(z_2, y),$$

$$A_1(2) = S^2(x) + k^2(2) S^2(z_1) - 2k(2) C(z_1, x),$$

$$\beta(2) = C(x, y) + k^2(2) C(z_1, z_2) - k(2) \{C(z_1, y) + C(x, z_2)\},$$

$$\Delta(2) = A_2(2) + \mu^2 \{b^2(2) A_1(2) - 2b(2) \beta(2)\}.$$

The optimum value of $k(2)$ is $k_0(2) = C(z_2, y) / S_2(z_2)$.

By substituting the optimum regression coefficient $b_0(2) = \beta(2) / A_1(2)$, obtained by minimizing the variance of $T_m(2)$, in (4.9) and assuming that $A_2(2) = A_1(2) = A(2)$ we have

$$V_0(T_2) = \frac{1}{n} \left[\frac{A^2(2) - \mu \beta^2(2)}{A^2(2) - \mu^2 \beta^2(2)} \right] A(2). \tag{4.10}$$

If the optimum μ is substituted in (4.10), the variance becomes

$$V_0(T_2) = \frac{1}{2n} [A(2) + \{A^2(2) - \beta^2(2)\}^{1/2}]. \tag{4.11}$$

4.3 Case (iii)

The third way of utilising available auxiliary information to improve the estimate of the current population total, Y , under the given sampling scheme is similar to the second. The only difference is that the population total of the auxiliary characteristic is not known; however, its FSU population mean is known for the selected FSUs.

This is given as

$$T_3 = \theta(3) T_m(3) + (1 - \theta(3)) T_u(3), \tag{4.12}$$

where

$$\begin{aligned} T_m(3) = & \frac{1}{\lambda n} \sum_{i=1}^{\lambda n} \frac{M_i}{P_i} \{ \bar{y}_i - k(3) (\bar{z}_{2i} - \bar{Z}_{2i}) \} \\ & - b(3) \left[\frac{1}{\lambda n} \sum_{i=1}^{\lambda n} \frac{M_i}{P_i} \{ \bar{x}_i - k(3) (\bar{z}_{1i} - \bar{Z}_{1i}) \} \right. \\ & \left. - \frac{1}{n} \sum_{i=1}^n \frac{M_i}{P_i} \{ \bar{x}_i - k(3) (\bar{z}_{1i} - \bar{Z}_{1i}) \} \right], \end{aligned}$$

and

$$T_u(3) = \frac{1}{n\mu} \sum_{i=1}^{n\mu} \frac{M_i}{P_i} \{ \bar{y}_i - k(3) (\bar{z}_{2i} - \bar{Z}_{2i}) \}.$$

For this estimator, we suppose that the values of both the main variate and the auxiliary variate are obtained for every SSU in the sample on both occasions. We also assume that the population mean, \bar{Z}_i , of the auxiliary variate is known for the selected FSUs.

The optimum variance of T_3 for a given λ is given as

$$V_0(T_3) = \frac{1}{n} [A_2(3) + \mu \{ b^2(3) A_1(3) - 2b(3) \beta(3) \}] A_2(3) / \Delta(3) \tag{4.13}$$

while the optimum weight is as usual obtained as

$$\theta_0(3) = \lambda A_2(3) / \Delta(3),$$

where

$$A_2(3) = S^2(y) + k^2(3) S_w^2(z_2) - 2k(3) C_w(z_2, y),$$

$$A_1(3) = S^2(x) + k^2(3) S_w^2(z_1) - 2k(3) C_w(z_1, x),$$

$$\beta(3) = C(x, y) + k^2(3) C_w(z_1, z_2) - k(3) \{C_w(z_1, y) + C_w(z_2, x)\},$$

$$\Delta(3) = A_2(3) + \mu^2 \{b^2(3) A_1(3) - 2b(3) \beta(3)\}.$$

The optimum value of $k(3)$ is $k_0(3) = C_w(z_2, y) / S_w^2(z_2)$.

If the optimum regression coefficient is substituted in (4.13), and it is assumed that population variances are the same on both occasions, then (4.13) works out as

$$V_0(T_3) = \frac{1}{n} \left[\frac{A^2(3) - \mu \beta^2(3)}{A^2(3) - \mu^2 \beta^2(3)} \right] A(3). \quad (4.14)$$

When the optimum μ is substituted in (4.14), the variance is

$$V_0(T_3) = \frac{1}{2n} [A(3) + \{A^2(3) - \beta^2(3)\}^{1/2}]. \quad (4.15)$$

4.4 Efficiency of the Proposed Estimators

The variances given in (4.7), (4.11) and (4.15) will be used to compare the efficiencies of T_1 , T_2 and T_3 with respect to

$$T_0 = \frac{1}{n} \sum_{i=1}^n \frac{M_i \bar{y}_i}{P_i}.$$

T_0 is the estimator for y when there is no partial matching of units and no auxiliary information used. In addition, the efficiency of T_0 compared to the usual partial matching estimator T' , which uses no auxiliary information, will be presented to assist in understanding the performance of the proposed estimators.

The usual partial matching estimator is defined as

$$T' = \theta' T'_m + (1 - \theta') T'_u, \quad (4.16)$$

where

$$T'_m = \frac{1}{\lambda n} \sum_{i=1}^{\lambda n} \frac{M_i \bar{y}_i}{P_i} - b' \left\{ \frac{1}{\lambda n} \sum_{i=1}^{\lambda n} \frac{M_i \bar{x}_i}{P_i} - \frac{1}{n} \sum_{i=1}^n \frac{M_i \bar{x}_i}{P_i} \right\},$$

and

$$T'_u = \frac{1}{n \mu} \sum_{i=1}^{n \mu} \frac{M_i \bar{y}_i}{P_i}.$$

The optimum variance of T' , obtained using the optimum value of b' , $b'_0 = C(x,y)/S^2(x)$, and assuming $S^2(y) = S^2(x)$ is

$$V_0(T') = \frac{1}{n} \left[\frac{S^2(y) - \mu C(x,y)}{S^2(y) - \mu^2 C(x,y)} \right] S^2(y). \tag{4.17}$$

Substituting the optimum value of μ in (4.17), the variance of T' becomes

$$V_0(T') = \frac{1}{2n} [S^2(y) + \{S^4(y) - C^2(x,y)\}^{1/2}]. \tag{4.18}$$

To calculate the efficiencies, the following assumptions about the correlation coefficients and the constant k were made:

$$\rho_b(x,z_2) = \rho_b(z_1,y) = \rho_b(z_1,z_2) = \rho_b;$$

$$\rho_w(x,z_2) = \rho_w(z_1,y) = \rho_w(z_1,z_2) = \rho_w;$$

$$k(1) = k(2) = k(3) = 1.$$

The efficiencies have been presented for only the positive values of ρ_b and ρ_w , and a set of values of

$$\delta = S^2_w(y)/S^2_b(y), R_b = S^2_b(z)/S^2_b(y) \text{ and } R_w = S^2_w(z)/S^2_b(y).$$

Looking at Table 2, we observe that none of the strategies T_1 , T_2 and T_3 (sampling design and estimator) is uniformly more efficient than strategy T_0 . The contrary is true of T' , which is always more efficient than T_0 ; at worst, its gain over T_0 is small (see Table 1).

The results in Tables 1 and 2 show T_1 is to be preferred to T' only when $R_b = 0.05$; and when $\rho_b = 0.8$ and $R_b = 0.5$.

Table 1
The Efficiency of T' with Respect to T_0

ρ_b	δ	$\rho_w = 0.2$	$\rho_w = 0.8$
0.2	0.05	1.01	1.01
	0.5	1.01	1.04
	5.0	1.01	1.17
0.8	0.05	1.22	1.25
	0.5	1.11	1.25
	5.0	1.02	1.25

T_2 is better than T' when:

- (i) $\rho_w = 0.2, R_b = R_w = 0.05$;
- (ii) $\rho_b = \rho_w = 0.8, R_b = R_w = 0.05, 0.5$;
- (iii) $\delta = 0.5, 5.0, R_w = R_b = 0.05, \rho_w = 0.5, \rho_b = 0.2$ and $\rho_w = 0.8$.

T_3 is generally more efficient than T' when:

- (i) $\delta = 5.0, \rho_w = 0.8$;
- (ii) $\delta = 0.5, \rho_w = 0.8$ and $R_w = 0.05, 0.5$.

The maximum gain in efficiency of T' over T_0 is 25% (see Table 1). In Table 2, the maximum gain of T_1 over T_0 is 155%, which occurs when $\rho_b = \rho_w = 0.8, \delta = 0.05, R_b = 0.5$. The maximum gain in efficiency of T_2 over T_0 is 172%; this happens when $\rho_b = \rho_w = 0.8, \delta = R_w = 0.05$. We also observe that when $\rho_b = \rho_w = 0.8, \delta = R_w = 5.0$, the maximum gain of T_3 over T_0 is 104%. It is therefore evident that the use of an auxiliary variate has tremendously improved the efficiency of partial matching of units.

If we now take the three strategies T_1, T_2 and T_3 , and compare them among themselves, we conclude that none of the strategies is uniformly better than the other, even though the maximum gain in efficiency of T_2 over T_0 is higher than that of T_1 , which in turn is higher than the maximum gain of T_3 . In general T_1 is superior to T_2 when $\rho_w = 0.2$, while T_2 is better than T_1 when $\rho_w = 0.8$. T_1 is preferred to T_3 when $\rho_b = 0.8, \rho_w = 0.2$ and $R_b = 0.05, 0.5$, and also when $\rho_b = \rho_w = 0.8$ and $\delta = R_b = 0.05$. Finally T_3 is better than T_2 when $\rho_w = 0.8, R_b = 5.0$, and when $\rho_b = \rho_w = 0.2$ with $R_b = 0.5, 5.0$.

5. APPLICATION

The proposed estimators were applied to a survey of teak plantations. The aim was to estimate the average height of teak trees using the girth as the auxiliary information.

Table 2
The Efficiency of T_1 , T_2 , and T_3 with Respect to T_0

		$\rho_w = 0.2$								
		$R_b = 0.05$			$R_b = 0.5$			$R_b = 5.0$		
ρ_b	δ	0.05	R_w 0.5	5.0	0.05	R_w 0.5	5.0	0.05	R_w 0.5	Strategy
0.2	0.05	1.04	1.04	1.04	0.83	0.83	0.83	0.20	0.20	T_1
		1.01	0.73	0.18	0.81	0.62	0.17	0.20	0.19	T_2
		0.98	0.71	0.18	0.98	0.71	0.18	0.98	0.71	T_3
	0.5	1.03	1.03	1.03	0.87	0.87	0.87	0.27	0.27	T_1
		1.04	0.85	0.26	0.88	0.74	0.25	0.27	0.25	T_2
		1.02	0.84	0.26	1.02	0.84	0.26	1.02	0.84	T_3
5.0	1.02	1.02	1.02	0.97	0.97	0.97	0.60	0.60	T_1	
	1.04	1.03	0.67	0.99	0.99	0.65	0.60	0.60	T_2	
	1.03	1.03	0.67	1.03	1.03	0.67	1.03	1.03	T_3	
0.8	0.05	1.62	1.62	1.62	2.53	2.53	2.53	0.45	0.45	T_1
		1.53	0.94	0.19	2.35	1.23	0.20	0.45	0.38	T_2
		1.16	0.77	0.18	1.16	0.77	0.18	1.16	0.77	T_3
	0.5	1.34	1.34	1.34	1.74	1.74	1.74	0.45	0.45	T_1
		1.34	1.03	0.27	1.76	1.28	0.29	0.54	0.48	T_2
		1.11	0.88	0.26	1.11	0.88	0.26	1.11	0.88	T_3
5.0	1.07	1.07	1.07	1.13	1.13	1.13	0.83	0.83	T_1	
	1.10	1.09	0.69	1.16	1.15	0.72	0.84	0.83	T_2	
	1.05	1.03	0.67	1.05	1.03	0.67	1.05	1.03	T_3	

		$\rho_w = 0.8$										
		$R_b = 0.05$				$R_b = 0.5$			$R_b = 5.0$			
ρ_b	δ	5.0	0.05	R_w 0.5	5.0	0.05	R_w 0.5	5.0	0.05	R_w 0.5	5.0	Strategy
0.2	0.05	0.20	1.05	1.05	1.05	0.83	0.83	0.83	0.20	0.20	0.20	T_1
		0.11	1.07	0.85	0.23	0.85	0.70	0.21	0.19	0.19	0.12	T_2
		0.18	1.04	0.83	0.23	1.04	0.83	0.23	1.04	0.83	0.23	T_3
	0.5	0.27	1.06	1.06	0.89	0.89	0.89	0.89	0.27	0.27	0.27	T_1
		0.15	1.21	1.30	0.41	1.00	1.06	0.38	0.28	0.28	0.19	T_2
		0.26	1.18	1.26	0.41	1.18	1.26	0.41	1.18	1.26	0.41	T_3
	5.0	0.60	1.17	1.17	1.17	1.09	1.09	1.09	0.62	0.62	0.62	T_1
		0.46	1.31	1.64	2.03	1.22	1.51	1.87	0.67	0.76	0.84	T_2
		0.67	1.30	1.63	2.00	1.30	1.63	2.00	1.30	1.63	2.00	T_3
0.8	0.05	0.45	1.65	1.65	1.65	2.55	2.55	2.55	0.46	0.46	0.46	T_1
		0.15	1.70	1.22	0.25	2.72	1.64	0.27	0.46	0.42	0.18	T_2
		0.18	1.27	0.98	0.24	1.26	0.98	0.24	1.27	0.98	0.24	T_3
	0.5	0.45	1.50	1.50	1.50	1.88	1.88	1.88	0.56	0.56	0.56	T_1
		0.21	1.75	1.83	0.46	2.34	2.65	0.50	0.59	0.61	0.31	T_2
		0.26	1.40	1.43	0.43	1.40	1.43	0.43	1.40	1.43	0.43	T_3
5.0	0.83	1.30	1.30	1.30	1.35	1.35	1.35	0.95	0.95	0.95	T_1	
	0.85	1.46	1.85	2.25	1.53	1.98	2.53	1.03	1.22	1.38	T_2	
	0.67	1.39	1.74	2.04	1.39	1.74	2.04	1.39	1.74	2.04	T_3	

Table 3
Estimated Efficiency of the Proposed Estimators with Respect to T_0 in the Estimation of the Average Height of Teak Trees

Estimators	Mean height (m)	Variance (m ²)	Estimated % Efficiency
T_0 (no matching)	20.04	6.3118	100
T' Partial matching	18.06	4.0680	155
T_1	17.86	0.0718	8791
T_2	17.31	0.0651	9635
T_3	17.99	4.0183	157

The teak trees used in this study were planted in 1965 with different spacings, producing plantations with the following number of trees per hectare: 2,000, 800, 400 and 250 trees. To measure the trees, an area of 40 metres by 40 metres was mapped out in each plantation after a sample of 8 plantations (FSUs) had been selected from 16 plantations, using the PPSWR scheme. The number of trees in each plantation was used as a measure of size. All the trees in the 40m by 40m area constituted the second stage units and the girth of each tree at breast height was measured. For the height measurements, a subsample of the trees was selected from the 40m by 40m area in each selected FSU. The first measurements were carried out in 1981 and the second in 1983. The sampling scheme used was the same as the one described in Section 2, with 50% matching of the FSUs.

The estimated efficiencies are given in Table 3. The sample estimates of the variance and covariance terms were used to obtain the optimum variances of T' , T_1 , T_2 and T_3 because the population values of these variances and covariances were not known. Therefore, the low values of the estimated optimum variances of T_1 and T_2 can be attributed partly to the nature of the sample data and partly to the nature of the estimators.

We observe that the estimator T_2 is more efficient than either T_1 or T_3 , while T_1 is more efficient than T_3 in the estimation of the average height of teak trees using the girth as the auxiliary information.

ACKNOWLEDGEMENTS

I am grateful to the referee and the associate editor for their useful comments for the improvement of this paper. I thank Dr. O. Abe of the Department of Statistics, University of Ibadan, Ibadan, for going over the draft of the revised paper.

REFERENCES

ABRAHAM, T.P., KHOSLA, R.K., and KATHURIA, O.P. (1969). Some investigations of the use of successive sampling in pest and disease surveys. *Journal of the Indian Society of Agricultural Statistics*, 21, 43 – 57.

COCHRAN, W.G. (1977). *Sampling Techniques*, (3rd. ed.). New York: John Wiley.

JESSEN, R.J. (1942). Statistical investigations of a sample survey for obtaining farm facts. *Iowa Agricultural Experimental Station Research Bulletin*, 304, 54-59.

KATHURIA, O.P. (1975). Some estimators in two-stage sampling on successive occasions with partial matching at both stages. *Sankhyā*, Ser. C, 37, 147 – 162.

- KATHURIA, O.P. (1978). Double sampling on successive occasions using a two-stage design. *Journal of the Indian Society of Agricultural Statistics*, 30, 49 – 64.
- KATHURIA, O.P., and SINGH, D. (1971). Relative efficiencies of some alternative procedures in two-stage sampling on successive occasions. *Journal of the Indian Society of Agricultural Statistics*, 23, 101 – 114.
- SINGH, S., and SRIVASTAVA, A.K. (1973). Use of auxiliary information in two-stage successive sampling. *Journal of the Indian Society of Agricultural Statistics*, 25, 101 – 114.
- SINGH, D. (1968). Estimates in successive sampling using a multistage design. *Journal of the American Statistical Association*, 63, 99 – 112.

'Some Optimality Results in the Presence of Nonresponse' by V.P. Godambe and M.E. Thompson, *Survey Methodology* (1986), 12, 29-36.

Formula (2.6), the definition of the optimum estimating function in $\bar{H}''(p, \mathbf{q})$, should be

$$h''^* = \sum_{i \in S'} (y_i - \theta x_i) \alpha_i / \pi_i q_i.$$

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 10, No. 2 and onward) of *Survey Methodology* as a guide and note particularly the following points:

1. Layout

- 1.1 Manuscripts should be typed on white bond paper of standard size ($8\frac{1}{2} \times 11$ inch), one side only, entirely double spaced with margins of at least $1\frac{1}{2}$ inches on all sides.
- 1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(\cdot)" and "log(\cdot)", etc.
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (e.g., w, ω ; o, O, 0; l, 1).
- 3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, priez d'examiner un numéro récent de Techniques d'enquête (à partir du vol. 10, n° 2) et de noter les points suivants:

1. **Présentation**
 - 1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8 1/2 par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1 1/2 pouce tout autour.
 - 1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés. Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
 - 1.4 Les remerciements doivent paraître à la fin du texte.
 - 1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.
2. **Résumé**

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. Rédaction

- 3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
- 3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(.) et log(.) etc.
- 3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
- 3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
- 3.5 Distinguer clairement les caractères ambigus (comme w, ω; o, O; 1, l).
- 3.6 Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.

4. Figures et tableaux

- 4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
- 4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois.)

5. Bibliographie

- 5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.
- 5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

'Résultats optimaux en situation de non-réponse' par V.P. Godambe et M.E. Thompson, Techniques d'enquête (1986), 12, 31-39.

L'équation (2.6), qui donne la définition de la fonction d'estimation optimale dans $H''(p, q)$, devrait être

$$h''_* = \sum_{l \in S'} (y_l - \theta x_l) \alpha_l / \pi_l q_l.$$

- JESSEN, R.J. (1942). Statistical investigations of a sample survey for obtaining farm facts. *Iowa Agricultural Experimental Station Research Bulletin*, 304, 54-59.
- KATHURIA, O.P. (1975). Some estimators in two-stage sampling on successive occasions with partial matching at both stages. *Sankhya, Sér. C*, 37, 147-162.
- KATHURIA, O.P. (1978). Double sampling on successive occasions using a two-stage design. *Journal of the Indian Society of Agricultural Statistics*, 30, 49-64.
- KATHURIA, O.P. et SINGH, D. (1971). Relative efficiencies of some alternative procedures in two-stage sampling on successive occasions. *Journal of the Indian Society of Agricultural Statistics*, 23, 101-114.
- SINGH, S., et SRIVASTAVA, A.K. (1973). Use of auxiliary information in two-stage successive sampling. *Journal of the Indian Society of Agricultural Statistics*, 25, 101-114.
- SINGH, D. (1968). Estimates in successive sampling using a multistage design. *Journal of the American Statistical Association*, 63, 99-112.

Tableau 3

Efficacité estimée des estimateurs proposés par rapport à T_0 dans l'estimation de la hauteur moyenne des arbres dans des plantations de teck				
Estimateurs	Hauteur moyenne (m)	Variance (m^2)	Efficacité estimée en %	
T_0 (sans appariement)	20.04	6.3118	100	
T' (appariement partiel)	18.06	4.0680	155	
T_1	17.86	0.0718	8791	
T_2	17.31	0.0651	9635	
T_3	17.99	4.0183	157	

Les arbres utilisés dans l'enquête ont été plantés en 1965 suivant différents espacements, ce qui a produit des plantations ayant le nombre d'arbres suivant par hectare: 2,000, 800, 400 et 250. Pour mesurer la hauteur des arbres, un périmètre de 40 mètres sur 40 a été tracé dans chacune des 8 plantations (USPD) prélevées parmi 16 plantations à l'aide d'un plan de sondage avec PPTAR. Le nombre d'arbres dans chaque plantation a été utilisé comme mesure de la taille. Tous les arbres à l'intérieur du périmètre de 40 m sur 40 formaient les unités de sondage du second degré et leur circonférence à hauteur de poitrine a été mesurée. Pour le calcul de la hauteur, un sous-échantillon d'arbres a été sélectionné des arbres de périmètre 40m sur 40 dans chaque USPD choisie. Une première série de calculs a été faite en 1981 et une seconde en 1983. Le plan de sondage utilisé était le même que celui qui a été décrit dans la section 2 et comportait un appariement partiel des USPD dans une propor-

tion de 50 %.

Les valeurs estimées de l'efficacité sont présentées dans le tableau 3. Les estimations de la variance et de la covariance de l'échantillon ont été utilisées pour calculer les variances optimums de T' , T_1 , T_2 et T_3 parce que les valeurs de ces variances et covariances pour l'ensemble de la population n'étaient pas connues. Par conséquent, le fait que les valeurs obtenues pour les variances estimées optimums de T_1 et T_2 sont faibles est attribuable d'une part, à l'utilisation de données d'échantillon et, d'autre part, à la nature même des estimateurs. Nous constatons que l'estimateur T_2 est plus efficace que T_1 et que T_3 , tandis que T_1 est plus efficace que T_3 dans l'estimation de la hauteur moyenne des arbres à l'aide de la cir-

conférence comme information auxiliaire.

REMERCIEMENTS

Je tiens à remercier l'arbitre et le rédacteur associé, dont les précieux commentaires m'ont aidé à améliorer le présent document. Je remercie également Dr. O. Abe du Département de statistique de l'Université d'Ibadan, pour les retouches apportées à la version préliminaire du document révisé.

BIBLIOGRAPHIE

- ABRAHAM, T.P., KHOSLA, R.K., et KATHURIA, O.P. (1969). Some investigations of the use of successive sampling in pest and disease surveys. *Journal of the Indian Society of Agricultural Statistics*, 21, 43-57.
- COCHRAN, W.G. (1977). *Sampling Techniques*, (3^e éd). New York: John Wiley.

Efficacité de T_1 , T_2 , et T_3 par rapport à T_0

p_b		δ	$R_b = 0.05$												$R_b = 0.5$												Stratè- gie																				
			$p_w = 0.8$												$R_b = 5.0$																																
0.8		0.5	0.05				5.0				0.5		5.0				0.05		5.0				R_w		0.05		R_w																				
T_3	T_2	T_1	0.45	1.50	1.75	1.40	1.30	1.46	1.39	1.30	1.43	1.50	1.50	1.85	1.83	0.26	0.18	0.45	1.65	1.70	1.27	0.98	0.24	1.26	2.55	2.72	1.64	0.98	1.88	2.65	1.43	1.98	1.74	1.35	1.30	1.43	0.95	1.22	1.74	2.04	1.39	1.03	0.95	0.95	1.38	2.04	T_3
0.67	0.85	0.83	0.21	0.45	0.85	0.26	0.67	0.60	0.46	0.67	1.30	1.65	1.70	1.05	0.20	0.18	0.27	1.06	1.21	1.18	1.26	1.63	1.17	1.64	2.03	1.22	1.09	1.09	1.51	1.63	2.55	2.72	1.64	0.98	1.88	2.65	1.43	1.98	1.74	1.35	1.30	1.43	0.95	1.22			
0.67	0.85	0.83	0.21	0.45	0.85	0.26	0.67	0.60	0.46	0.67	1.30	1.65	1.70	1.05	0.20	0.18	0.27	1.06	1.21	1.18	1.26	1.63	1.17	1.64	2.03	1.22	1.09	1.09	1.51	1.63	2.55	2.72	1.64	0.98	1.88	2.65	1.43	1.98	1.74	1.35	1.30	1.43	0.95	1.22			
0.67	0.85	0.83	0.21	0.45	0.85	0.26	0.67	0.60	0.46	0.67	1.30	1.65	1.70	1.05	0.20	0.18	0.27	1.06	1.21	1.18	1.26	1.63	1.17	1.64	2.03	1.22	1.09	1.09	1.51	1.63	2.55	2.72	1.64	0.98	1.88	2.65	1.43	1.98	1.74	1.35	1.30	1.43	0.95	1.22			
0.67	0.85	0.83	0.21	0.45	0.85	0.26	0.67	0.60	0.46	0.67	1.30	1.65	1.70	1.05	0.20	0.18	0.27	1.06	1.21	1.18	1.26	1.63	1.17	1.64	2.03	1.22	1.09	1.09	1.51	1.63	2.55	2.72	1.64	0.98	1.88	2.65	1.43	1.98	1.74	1.35	1.30	1.43	0.95	1.22			
0.67	0.85	0.83	0.21	0.45	0.85	0.26	0.67	0.60	0.46	0.67	1.30	1.65	1.70	1.05	0.20	0.18	0.27	1.06	1.21	1.18	1.26	1.63	1.17	1.64	2.03	1.22	1.09	1.09	1.51	1.63	2.55	2.72	1.64	0.98	1.88	2.65	1.43	1.98	1.74	1.35	1.30	1.43	0.95	1.22			
0.67	0.85	0.83	0.21	0.45	0.85	0.26	0.67	0.60	0.46	0.67	1.30	1.65	1.70	1.05	0.20	0.18	0.27	1.06	1.21	1.18	1.26	1.63	1.17	1.64	2.03	1.22	1.09	1.09	1.51	1.63	2.55	2.72	1.64	0.98	1.88	2.65	1.43	1.98	1.74	1.35	1.30	1.43	0.95	1.22			
0.67	0.85	0.83	0.21	0.45	0.85	0.26	0.67	0.60	0.46	0.67	1.30	1.65	1.70	1.05	0.20	0.18	0.27	1.06	1.21	1.18	1.26	1.63	1.17	1.64	2.03	1.22	1.09	1.09	1.51	1.63	2.55	2.72	1.64	0.98	1.88	2.65	1.43	1.98	1.74	1.35	1.30	1.43	0.95	1.22			
0.67	0.85	0.83	0.21	0.45	0.85	0.26	0.67	0.60	0.46	0.67	1.30	1.65	1.70	1.05	0.20	0.18	0.27	1.06	1.21	1.18	1.26	1.63	1.17	1.64	2.03	1.22	1.09	1.09	1.51	1.63	2.55	2.72	1.64	0.98	1.88	2.65	1.43	1.98	1.74	1.35	1.30	1.43	0.95	1.22			
0.67	0.85	0.83	0.21	0.45	0.85	0.26	0.67	0.60	0.46	0.67	1.30	1.65	1.70	1.05	0.20	0.18	0.27	1.06	1.21	1.18	1.26	1.63	1.17	1.64	2.03	1.22	1.09	1.09	1.51	1.63	2.55	2.72	1.64	0.98	1.88	2.65	1.43	1.98	1.74	1.35	1.30	1.43	0.95	1.22			
0.67	0.85	0.83	0.21	0.45	0.85	0.26	0.67	0.60	0.46	0.67	1.30	1.65	1.70	1.05	0.20	0.18	0.27	1.06	1.21	1.18	1.26	1.63	1.17	1.64	2.03	1.22	1.09	1.09	1.51	1.63	2.55	2.72	1.64	0.98	1.88	2.65	1.43	1.98	1.74	1.35	1.30	1.43	0.95	1.22			
0.67	0.85	0.83	0.21	0.45	0.85	0.26	0.67	0.60	0.46	0.67	1.30	1.65	1.70	1.05	0.20	0.18	0.27	1.06	1.21	1.18	1.26	1.63	1.17	1.64	2.03	1.22	1.09	1.09	1.51	1.63	2.55	2.72	1.64	0.98	1.88	2.65	1.43	1.98	1.74	1.35	1.30	1.43	0.95	1.22			
0.67	0.85	0.83	0.21	0.45	0.85	0.26	0.67	0.60	0.46	0.67	1.30	1.65	1.70	1.05	0.20	0.18	0.27	1.06	1.21	1.18	1.26	1.63	1.17	1.64	2.03	1.22	1.09	1.09	1.51	1.63	2.55	2.72	1.64	0.98	1.88	2.65	1.43	1.98	1.74	1.35	1.30	1.43	0.95	1.22			
0.67	0.85	0.83	0.21	0.45	0.85	0.26	0.67	0.60	0.46	0.67	1.30	1.65	1.70	1.05	0.20	0.18	0.27	1.06	1.21	1.18	1.26	1.63	1.17	1.64	2.03	1.22	1.09	1.09	1.51	1.63	2.55	2.72	1.64	0.98	1.88	2.65	1.43	1.98	1.74	1.35	1.30	1.43	0.95	1.22			
0.67	0.85	0.83	0.21	0.45	0.85	0.26	0.67	0.60	0.46	0.67	1.30	1.65	1.70	1.05	0.20	0.18	0.27	1.06	1.21	1.18	1.26	1.63	1.17	1.64	2.03	1.22	1.09	1.09	1.51	1.63	2.55	2.72	1.64	0.98	1.88	2.65	1.43	1.98	1.74	1.35	1.30	1.43	0.95	1.22			
0.67	0.85	0.83	0.21	0.45	0.85	0.26	0.67	0.60	0.46	0.67	1.30	1.65	1.70	1.05	0.20	0.18	0.27	1.06	1.21	1.18	1.26	1.63	1.17	1.64	2.03	1.22	1.09	1.09	1.51	1.63	2.55	2.72	1.64	0.98	1.88	2.65	1.43	1.98	1.74	1.35	1.30	1.43	0.95	1.22			
0.67	0.85	0.83	0.21	0.45	0.85	0.26	0.67	0.60	0.46	0.67	1.30	1.65	1.70	1.05	0.20	0.18	0.27	1.06	1.21	1.18	1.26	1.63	1.17	1.64	2.03	1.22	1.09	1.09	1.51	1.63	2.55	2.72	1.64	0.98	1.88	2.65	1.43	1.98	1.74	1.35	1.30	1.43	0.95	1.22			
0.67	0.85	0.83	0.21	0.45	0.85	0.26	0.67	0.60	0.46	0.67	1.30	1.65	1.70	1.05	0.20	0.18	0.27	1.06	1.21	1.18	1.26	1.63	1.17	1.64	2.03	1.22	1.09	1.09	1.51	1.63	2.55	2.72	1.64	0.98	1.88	2.65	1.43	1.98	1.74	1.35	1.30	1.43	0.95	1.22			
0.67	0.85	0.83	0.21	0.45	0.85	0.26	0.67	0.60	0.46	0.67	1.30	1.65	1.70	1.05	0.20	0.18	0.27	1.06	1.21	1.18	1.26	1.63	1.17	1.64	2.03	1.22	1.09	1.09	1.51	1.63	2.55	2.72	1.64	0.98	1.88	2.65	1.43	1.98	1.74	1.35	1.30	1.43	0.95	1.22			
0.67	0.85	0.83	0.21	0.45	0.85	0.26	0.67	0.60	0.46	0.67	1.30	1.65	1.70	1.05	0.20	0.18	0.27	1.06	1.21	1.18	1.26	1.63	1.17	1.64	2.03	1.22	1.09	1.09	1.51	1.63	2.55	2.72	1.64	0.98	1.88	2.65	1.43	1.98	1.74	1.35	1.30	1.43	0.95	1.22			
0.67	0.85	0.83	0.21	0.45	0.85	0.26	0.67	0.60	0.46	0.67	1.30	1.65	1.70	1.05	0.20	0.18	0.27	1.06	1.21	1.18	1.26	1.63	1.17	1.64	2.03	1.22	1.09	1.09	1.51	1.63	2.55	2.72	1.64	0.98	1.88	2.65	1.43	1.98	1.74	1.35	1.30	1.43	0.95	1.22			
0.67	0.85	0.83	0.21	0.45	0.85	0.26	0.67	0.60	0.46	0.67	1.30	1.65	1.70	1.05	0.20	0.18	0.27	1.06	1.21	1.18	1.26	1.63	1.17	1.64	2.03	1.22	1.09	1.09	1.51	1.63	2.55	2.72	1.64	0.98	1.88	2.65	1.43	1.98	1.74	1.35	1.30	1.43	0.95	1.22			
0.67	0.85	0.83	0.21	0.45	0.85	0.26	0.67	0.60	0.46	0.67	1.30	1.65	1.70	1.05	0.20	0.18	0.27	1.06	1.21	1.18	1.26	1.63	1.17	1.64	2.03	1.22	1.09	1.09	1.51	1.63	2.55	2.72	1.64	0.98	1.88	2.65	1.43	1.98	1.74	1.35	1.30	1.43	0.95	1.22			
0.67	0.85	0.83	0.21	0.45	0.85	0.26	0.67	0.60	0.46	0.67	1.30	1.65	1.70	1.05	0.20	0.18	0.27	1.06	1.21	1.18	1.26	1.63	1.17	1.64	2.03	1.22	1.09	1.09	1.51	1.63	2.55	2.72	1.64	0.98	1.88	2.65	1.43	1.98	1.74	1.35	1.30	1.43	0.95	1.22			
0.67	0.85	0.83	0.21	0.45	0.85	0.26	0.67	0.60	0.46	0.67	1.30	1.65	1.70	1.05	0.20	0.18	0.27	1.06	1.21	1.18	1.26	1.63	1.17	1.64	2.03	1.22	1.09	1.09	1.51	1.63	2.55	2.72	1.64	0.98	1.88	2.65	1.43	1.98	1.74	1.35	1.30	1.43	0.95	1.22			
0.67	0.85	0.83	0.21	0.45	0.85	0.26	0.67	0.60	0.46	0.67	1.30	1.65	1.70	1.05	0.20	0.18	0.27	1.06	1.21	1.18	1.26	1.63	1.17	1.64	2.03	1.22	1.09	1.09	1.51	1.63	2.55	2.72	1.64	0.98	1.88	2.65	1.43	1.98	1.74	1.35	1.30	1.43	0.95	1.22			
0.67	0.85	0.83	0.21	0.45	0.85	0.26	0.67	0.60	0.46	0.67	1.30	1.65	1.70	1.05	0.20	0.18	0.27	1.06	1.21	1.18	1.26	1.63	1.17	1.64	2.03	1.22	1.09	1.09	1.51	1.63	2.55	2.72	1.64	0.98	1.88	2.65	1.43	1.98	1.74	1.35	1.30	1.43	0.95	1.22			
0.67	0.85	0.83	0.21	0.45	0.85	0.26	0.67	0.60	0.46	0.67	1.30	1.65	1.70	1.05	0.20	0.18	0.27	1.06	1.21	1.18	1.26	1.63	1.17	1.64	2.03	1.22	1.09	1.09	1.51	1.63	2.55	2.72	1.64	0.98	1.88	2.65	1.43	1.98	1.74	1.35	1.30	1.43	0.95	1.22			
0.67	0.85	0.83	0.21	0.45	0.85	0.26	0.67	0.60	0.46	0.67	1.30	1.65	1.70	1.05	0.20	0.18	0.27	1.06	1.21	1.18	1.26	1.63	1.17	1.64	2.03	1.22	1.09	1.09	1.51	1.63	2.55	2.72	1.64	0.98	1.88	2.65	1.43	1.98	1.74	1.35	1.30	1.43	0.95	1.22			
0.67	0.85	0.83	0.21	0.45	0.85	0.26	0.67	0.60	0.46	0.67	1.30	1.65	1.70	1.05	0.20	0.18	0.27	1.06	1.21	1.18	1.26	1.63	1.1																								

p_b	δ	$p_w = 0.2$	$p_w = 0.8$
0.2	0.05	1.01	1.01
	0.5	1.01	1.04
	5.0	1.01	1.17
0.8	0.05	1.22	1.25
	0.5	1.11	1.25
	5.0	1.02	1.25

Efficacité de T' par rapport à T_0

Tableau 1

T_2 est meilleur que T' quand

(i) $p_w = 0.2, R_b = R_w = 0.05;$

(ii) $p_b = p_w = 0.8, R_b = R_w = 0.05, 0.5;$

(iii) $\delta = 0.5, 5.0, R_w = R_b = 0.05, 0.05, p_b = 0.2$ et $p_w = 0.8.$

T_3 est généralement plus efficace que T' quand

(i) $\delta = 5.0, p_w = 0.8;$

(ii) $\delta = 0.5, p_w = 0.8$ and $R_w = 0.05, 0.5.$

Le gain maximum d'efficacité de T' par rapport à T_0 est de 25% (voir tableau 1). D'après les chiffres du tableau 2, le gain maximum de T_1 par rapport à T_0 est de 155%; il est obtenu quand $p_b = p_w = 0.8, \delta = 0.05$ et $R_b = 0.5$. Le gain maximum d'efficacité de T_2 par rapport à T_0 est de 172%; il est obtenu quand $p_b = p_w = 0.8$ et $\delta = R_w = 0.05$. Nous constatons également que lorsque $p_b = p_w = 0.8$, et $\delta = R_w = 5.0$, le gain maximum de T_3 par rapport à T_0 est de 104%. Il est donc évident que l'utilisation d'une variable auxiliaire a beaucoup amélioré l'efficacité de l'appariement partiel des unités.

Si maintenant nous comparons entre elles les trois stratégies T_1, T_2 et T_3 , nous pouvons conclure qu'aucune n'est uniformément meilleure qu'une autre, même si le gain maximum d'efficacité de T_2 est supérieur au gain maximum d'efficacité de T_1 , lui-même plus élevé que le gain maximum de T_3 par rapport à T_0 . En général, T_1 est meilleur que T_2 quand $p_w = 0.2$, tandis que T_2 est meilleur que T_1 quand $p_w = 0.8, T_1$ est préférable à T_3 quand $p_b = 0.8, p_w = 0.2$ et $R_b = 0.05, 0.5$ ou quand $p_b = p_w = 0.8$ et $\delta = R_b = 0.05$. Enfin, T_3 est meilleur que T_2 quand $p_w = 0.8$ et $R_b = 5.0$ ou quand $p_b = p_w = 0.2$ et $R_b = 0.5, 5.0$.

5. APPLICATION

Les estimateurs proposés ont été appliqués à une enquête sur la hauteur des arbres dans des plantations de teck. L'objectif était d'estimer la hauteur moyenne des arbres en utilisant la circonférence des troncs comme information auxiliaire.

où

$$T'_m = \frac{1}{\lambda n} \sum_{i=1}^n \frac{P_i}{M_{iY_i}} - b', \left\{ \frac{1}{\lambda n} \sum_{i=1}^n \frac{P_i}{M_{iX_i}} - \frac{1}{n} \sum_{i=1}^n \frac{P_i}{M_{iX_i}} \right\},$$

et

$$T'_n = \frac{1}{n\mu} \sum_{i=1}^n \frac{P_i}{M_{iY_i}}.$$

La variance optimum de T' , obtenue en utilisant la valeur optimum de b' , $b'_0 = C(x,y) / S^2(x)$, et en supposant que $S^2(y) = S^2(x)$ est

$$V_0(T') = \frac{1}{n} \left[\frac{S^2(y) - \mu C(x,y)}{S^2(x,y)} \right] S^2(y). \tag{4.17}$$

Si on substitue la valeur optimum de μ dans (4.17), la variance de T' devient

$$V_0(T') = \frac{1}{2n} [S^2(y) + \{S^4(y) - C^2(x,y)\}^{1/2}]. \tag{4.18}$$

Pour calculer l'efficacité des divers estimateurs, les hypothèses suivantes ont été faites au sujet des coefficients de corrélation et de la constante k :

$$\rho_b(x,z_2) = \rho_b(z_1,y) = \rho_b(z_1,z_2) = \rho_b;$$

$$\rho_w(x,z_2) = \rho_w(z_1,y) = \rho_w(z_1,z_2) = \rho_w;$$

$$k(1) = k(2) = k(3) = 1.$$

Les valeurs de l'efficacité n'ont été présentées que pour des valeurs positives de ρ_b et ρ_w et une série de valeurs de

$$\delta = S^2_w(y) / S^2_b(y), R_b = S^2_b(z) / S^2_b(y) \text{ et } R_w = S^2_w(z) / S^2_b(y).$$

Si on regarde le tableau 2, on constate qu'aucune des stratégies T_1 , T_2 ou T_3 (plan de sondage et estimateur) n'est uniformément plus efficace que la stratégie T_0 . C'est le contraire pour T' , qui est toujours plus efficace que T_0 , et qui, au pire, n'offre qu'un faible gain par rapport à T_0 (voir tableau 1).
D'après les résultats des tableaux 1 et 2, il faut préférer T_1 à T' seulement lorsque $R_b = 0.05$, $\rho_b = 0.8$ et $R_b = 0.5$.

tandis que le poids optimum est donné, comme d'habitude, par l'expression suivante:

$$\theta_0(3) = \lambda A_2(3) / \Delta(3),$$

où

$$A_2(3) = S^2(y) + k^2(3) S^2_w(z_2) - 2k(3) C_w(z_2, y),$$

$$A_1(3) = S^2(x) + k^2(3) S^2_w(z_1) - 2k(3) C_w(z_1, x),$$

$$\beta(3) = C(x, y) + k^2(3) C_w(z_1, z_2) - k(3) \{ C_w(z_1, y) + C_w(z_2, x) \},$$

$$\Delta(3) = A_2(3) + \mu^2 \{ b^2(3) A_1(3) - 2b(3) \beta(3) \}.$$

La valeur optimum de $k(3)$ est $k_0(3) = C_w(z_2, y) / S^2_w(z_2)$.

Si on substitue le coefficient optimum de régression dans (4.13) et si on suppose que la variance de population est la même dans les deux périodes, (4.13) se ramène alors à

$$V_0(T_3) = \frac{1}{n} \left[\frac{A_2(3) - \mu \beta^2(3)}{A_2(3)} \right] A(3). \tag{4.14}$$

Quand la valeur optimum de μ est substituée dans (4.14), la variance devient

$$V_0(T_3) = \frac{1}{n} [A(3) + \{A^2_2(3) - \beta^2(3)\} \frac{1}{2}]. \tag{4.15}$$

4.4 Efficacité des estimateurs proposés

Nous utiliserons les variances données en (4.7), (4.11) et (4.15) pour comparer l'efficacité des trois estimateurs, T_1 , T_2 et T_3 par rapport à

$$T_0 = \frac{1}{n} \sum_{i=1}^n \frac{M_i Y_i}{P_i}.$$

T_0 est l'estimateur de y quand il n'y a pas d'appariement partiel des unités et qu'aucune information auxiliaire n'est utilisée. Nous avons aussi comparé l'efficacité de T_0 par rapport à l'estimateur habituel avec appariement partiel, T' , qui n'utilise pas d'information auxiliaire, pour mieux faire comprendre la performance des estimateurs proposés. L'estimation habituelle avec appariement partiel est définie comme suit:

$$T' = \theta' T'_m + (1 - \theta') T'_n,$$

(4.16)

La valeur optimum de $k(2)$ est $k_0(2) = C(z_2, y) / S_2(z_2)$.

En substituant le coefficient optimum de régression, $b_0(2) = \beta(2) / A_1(2)$, obtenue en minimisant la variance de $T_m(2)$, dans (4.9) et en supposant que $A_2(2) = A_1(2) = A(2)$, on obtient

$$V_0(T_2) = \frac{1}{n} \left[\frac{A_2(2) - \mu\beta^2(2)}{A_2(2) - \mu^2\beta^2(2)} \right] A(2). \quad (4.10)$$

Si la valeur optimum de μ est substituée dans (4.10), la variance devient

$$V_0(T_2) = \frac{1}{2n} [A(2) + \{A_2(2) - \beta^2(2)\}]. \quad (4.11)$$

4.3 Cas (iii)

La troisième façon d'utiliser l'information auxiliaire connue pour améliorer l'estimation du total de population de la période donnée, Y , dans le cadre d'un plan de sondage donné ressemble beaucoup à la deuxième façon. La seule différence est qu'on ne connaît pas le total de population de la caractéristique auxiliaire; par contre, on connaît la moyenne de population pour les USPD choisis.

L'estimateur s'exprime ainsi:

$$T_3 = \theta(3) T_m(3) + (1 - \theta(3)) T_n(3), \quad (4.12)$$

ou

$$T_m(3) = \frac{1}{\lambda n} \sum_{i=1}^{\lambda n} \frac{P_i}{M_i} \{Y_i - k(3) (\bar{z}_{2i} - \bar{Z}_{2i})\} \\ - b(3) \left[\frac{1}{\lambda n} \sum_{i=1}^{\lambda n} \frac{P_i}{M_i} \{X_i - k(3) (\bar{z}_{1i} - \bar{Z}_{1i})\} \right],$$

et

$$T_n(3) = \frac{1}{n\mu} \sum_{i=1}^{n\mu} \frac{P_i}{M_i} \{Y_i - k(3) (\bar{z}_{2i} - \bar{Z}_{2i})\}.$$

Pour cet estimateur, on suppose que les valeurs tant de la variable principale que de la variable auxiliaire sont obtenues pour chaque USPD de l'échantillon dans les deux périodes. On suppose également que la moyenne de population, \bar{Z}_i , de la variable auxiliaire est connue pour les USPD choisis.

La variance optimum de T_3 pour une valeur donnée de λ s'exprime ainsi:

$$V_0(T_3) = \frac{1}{n} [A_2(3) + \mu \{b^2(3) A_1(3) - 2b(3) \beta(3)\} A_2(3) / \Delta(3)] \quad (4.13)$$

4.2 Cas (iii)

Le deuxième estimateur est l'estimateur habituel dans lequel l'information relative aussi bien à la caractéristique principale qu'à la caractéristique auxiliaire a été obtenue des unités déclarantes et dans lequel le total de population de la caractéristique auxiliaire est connu. Il s'exprime ainsi:

(4.8) $T_2 = \theta(2) T_m(2) + (1 - \theta(2)) T_n(2),$

où

$$T_m(2) = \frac{1}{\lambda_n} \sum_{i=1}^t \left\{ \frac{M_{iy_i}}{P_i} - k(2) \right\} \left(\frac{M_{iz_{2i}}}{P_i} - Z_2 \right) - b(2) \left[\frac{1}{\lambda_n} \sum_{i=1}^t \left\{ \frac{M_{ix_i}}{P_i} - k(2) \right\} \left(\frac{M_{iz_{1i}}}{P_i} - Z_1 \right) \right],$$

et

$$T_n(2) = \frac{1}{n} \sum_{i=1}^t \left\{ \frac{M_{ix_i}}{P_i} - k(2) \right\} \left(\frac{M_{iz_{2i}}}{P_i} - Z_2 \right).$$

Ici, le total de population global de la variable auxiliaire est connu dans les deux périodes. En outre, l'information sur la variable auxiliaire, z_{ij} , est obtenue pour chaque USSD de l'échantillon. Il s'agit de la façon habituelle d'utiliser l'information auxiliaire dans les plans de sondage décrits dans les ouvrages portant sur le sujet. On peut montrer que la variance optimum de T_2 est

(4.9) $V_0(T_2) = \frac{1}{n} [A_2(2) + \mu \{b^2(2)A_1(2) - 2b(2)\beta(2)\}] A_2(2) / \Delta(2)$

et que le poids optimum est

$$\theta_0(2) = \lambda A_2(2) / \Delta(2)$$

où

$$A_2(2) = S_2^2(y) + k^2(2) S_2^2(z_2) - 2k(2) C(z_2, y),$$

$$A_1(2) = S_2^2(x) + k^2(2) S_2^2(z_1) - 2k(2) C(z_1, x),$$

$$\beta(2) = C(x, y) + k^2(2) C(z_1, z_2) - k(2) \{ C(z_1, y) + C(x, z_2) \},$$

$$\Delta(2) = A_2(2) + \mu^2 \{ b^2(2) A_1(2) - 2b(2) \beta(2) \}.$$

On peut montrer qu'en utilisant la méthode proposée par Jessen (1942), la valeur optimum de $V(T_1)$ pour une valeur donnée de λ est

(4.2)
$$V_0(T_1) = \frac{n}{1} [A_2(1) + \mu \{b^2(1)A_1(1) - 2b(1)\beta(1)\}] A_2(1) / \Delta(1)$$

où

$$A_1(1) = S^2(x) + k^2(1) S_b^2(z_1) - 2k(1) C_b(z_1, x),$$

$$\beta(1) = C(x, y) + k^2(1) C_b(z_1, z_2) - k(1) \{C_b(x, z_2) + C_b(z_1, y)\},$$

$$\Delta(1) = A_2(1) + \mu^2 \{b^2(1) A_1(1) - 2b(1) \beta(1)\}.$$

Lorsqu'on minimise la variance de $T_m(1)$, la valeur optimum de $b(1)$ est

$$b_0(1) = \beta(1) / A_1(1).$$

Si on substitue $b_0(1)$ dans (4.2), la variance optimum devient

(4.3)
$$V_0(T_1) = \frac{n}{1} \left[\frac{A_1(1) A_2(1) - \mu \beta^2(1)}{A_1(1) A_2(1) - \mu^2 \beta^2(1)} \right] A_2(1).$$

Lorsqu'on minimise $V_0(T_1)$ dans (4.2) par rapport à μ , la fraction d'appariement optimum se ramène à $\lambda_0 = 1 - \mu_0$ où

(4.4)
$$\mu_0 = A_2(1) [A_2(1) + \{A_2^2(1) + A_2(1) (b^2(1) A_1(1) - 2b(1) \beta(1))\}^{-1/2}]^{-1}.$$

Si $A_2(1) = A_1(1)$, c'est-à-dire si la variation de la population est la même dans les deux périodes, l'expression (4.3) donne

(4.5)
$$V_0(T_1) = \frac{n}{1} \left[\frac{A_2^2(1) - \mu \beta^2(1)}{A_2^2(1) - \mu^2 \beta^2(1)} \right] A(1)$$

et si on substitue $b_0(1)$ à $b(1)$, la fraction d'appariement optimum donnée dans l'équation (4.4), μ_0 , devient

(4.6)
$$\mu_0 = A(1) [A(1) + \{A^2(1) - \beta^2(1)\}^{-1/2}]^{-1}.$$

Quand on substitue μ_0 dans (4.5), la variance se ramène à

(4.7)
$$V_0(T_1) = \frac{1}{2n} [A(1) + \{A^2(1) - \beta^2(1)\}^{-1/2}].$$

4. ESTIMATEURS DE TOTAUX DE POPULATION ET VARIANCES OPTIMUMS

4.1 Cas (i)

Le premier estimateur du total de population Y , de la seconde période est utilisé lorsqu'on n'a pas d'information sur la variable auxiliaire, mais qu'on connaît le total de population de la variable auxiliaire pour les USPD choisies. Il s'exprime ainsi:

$$T_1 = \theta(1) T_m(1) + (1 - \theta(1)) T_n(1) \quad (4.1)$$

où $\theta(1)$ est une constante choisie e telle sorte que la variance de T_1 , $V(T_1)$ est minimum, tandis que

$$T_m(1) = \frac{1}{\lambda n} \sum_{i=1}^l \left\{ \frac{M_{iY}}{P_i} - k(1) \left(\frac{Z_{2i}}{P_i} - Z_2 \right) \right\} \left[\frac{1}{\lambda n} \sum_{i=1}^l \left\{ \frac{M_{iX}}{P_i} - k(1) \left(\frac{Z_{1i}}{P_i} - Z_1 \right) \right\} \right] - b(1) \left[\frac{1}{\lambda n} \sum_{i=1}^l \left\{ \frac{M_{iX}}{P_i} - k(1) \left(\frac{Z_{1i}}{P_i} - Z_1 \right) \right\} \right]$$

est l'estimateur par différence de Y fondé sur l'échantillon apparié, que

$$T_n(1) = \frac{1}{n} \sum_{i=1}^l \left\{ \frac{M_{iY}}{P_i} - k(1) \left(\frac{Z_{2i}}{P_i} - Z_2 \right) \right\}$$

est l'estimateur de Y fondé sur l'échantillon non apparié et que $k(1)$ et $b(1)$ sont des constantes connues.

Pour cet estimateur, on suppose qu'on connaît le total de population de la variable auxiliaire, Z_i , pour chaque USPD choisie à la première période. On connaît également le total de population pour l'ensemble de USPD, Z , à chaque période. Aucune autre information sur la variable auxiliaire n'est obtenue des répondants ou des unités déclarantes (UD). Maintenant, en minimisant $V(T_1)$ par rapport à $\theta(1)$ et en résolvant l'équation, on obtient la valeur optimum suivante de $\theta(1)$

$$\theta_0(1) = \lambda A_2(1) / \Delta(1)$$

où

$$A_2(1) = S^2(y) + k^2(1) S_b^2(z_2) - 2k(1) C_b(z_2, y),$$

$$\Delta(1) = A_2(1) + \mu^2 \{ b^2(1) A_1(1) - 2b(1) \beta(1) \}.$$

La valeur optimum de $k(1)$ est obtenue en minimisant $V(T_n(1))$ par rapport à $k(1)$. Cela donne $k_0(1) = C_b(z_2, y) / S_b^2(z_2)$.

Cochran (1977, p. 306), qui stipule que si la i -ième USSD de S_1 est choisie θ_i fois ($i = 1, 2, \dots, n$), on tire θ_i sous-échantillons indépendants de taille m_i à partir des M_i USSD.

À la deuxième période, nous prélevons un échantillon de λn USSD ($0 < \lambda < 1$) à partir de S_1 selon un plan de sondage aléatoire simple et sans remise (SASSR). Les USSD choisies à la première période sont retenues pour chacune de ces λn USSD apparées. Ensuite, un nouvel échantillon de μn ($\mu = 1 - \lambda$) USSD est tiré indépendamment des n USSD par sondage avec PPTSR, avec P_i comme mesure normalisée de la taille de la i -ième USSD. Dans chacune des μn USSD, les USSD sont choisies de la même façon qu'à la première période.

3. NOTATION

Nous définissons $y_{ij}(x_{ij})$ comme la valeur de la variable à l'étude pour la j -ième USSD dans la i -ième période donnée (ou la période précédente). De plus, z_{hij} est définie comme la valeur de la variable auxiliaire pour la j -ième USSD dans la i -ième USSD à la h -ième période ($h = 1, 2$). Les moyennes de l'échantillon des USSD dans la i -ième USSD sont

$$x_i = \frac{1}{m_i} \sum_{m_i}^j x_{ij}, \quad y_i = \frac{1}{m_i} \sum_{m_i}^j y_{ij} \quad \text{et} \quad z_{hi} = \frac{1}{m_i} \sum_{m_i}^j z_{hij}.$$

Le total de population pour la i -ième USSD et le total de population pour l'ensemble des USSD qui correspondent à la variable auxiliaire sont

$$Z_{hi} = \sum_{M_i}^j z_{hij} \quad \text{et} \quad Z_h = \sum_N^j Z_{hi}.$$

Nous définissons aussi quelques autres notations de la façon suivante:

$$S_i^b(y) = \sum_N^i P_i \left(\frac{P_i}{Y} - Y \right)^2 \quad \text{est la variance entre les USSD;}$$

$$S_w^2(y) = \sum_N^i M_i^2 \left(\frac{P_i}{1} - \frac{m_i}{1} \right) S_{wi}^2(y) \quad \text{est la variance entre les USSD de l'ensemble des USSD;}$$

$$S_{wi}^2(y) = \frac{1}{M_i} \sum_{M_i}^j (y_{ij} - \bar{y}_i)^2 \quad \text{est la variance entre les USSD de la } i\text{-ième USSD;}$$

$$S_i^2(y) = S_i^b(y) + S_{wi}^2(y);$$

$$C_b(x, y) = \rho_b S_i^b(x) S_i^b(y) \quad \text{est la covariance de } x \text{ et } y \text{ entre les USSD;}$$

$$C_w(x, y) = \rho_w S_w(x) S_w(y) \quad \text{est la covariance de } x \text{ et } y \text{ entre les USSD de l'ensemble des USSD;}$$

$$C(x, y) = C_b(x, y) + C_w(x, y).$$

Les coefficients de corrélation entre x et y calculés entre les USSD et à l'intérieur des USSD sont respectivement ρ_b et ρ_w .

Quand une variable auxiliaire est très corrélée à la caractéristique étudiée, on peut améliorer l'estimation de la moyenne (ou du total) de population de cette caractéristique en utilisant la variable auxiliaire. Shvitar Singh et Srivastava (1973) ont utilisé de l'information auxiliaire pour améliorer l'estimateur de Singh (1968). Ils ont obtenu un estimateur linéaire non biaisé de la moyenne de population de la période la plus récente en utilisant des variables auxiliaires dont les moyennes de population étaient connues et ne changeaient pas d'une période à l'autre. Kathuria (1978) a poussé davantage dans cette voie en supposant que la moyenne de population de la variable auxiliaire n'est pas connue. Il a utilisé une technique d'échantillonnage double (ou sondage à deux phases) pour estimer d'abord la moyenne de population de la variable auxiliaire et ensuite la moyenne de population de la caractéristique à l'étude. Dans leurs ouvrages, Shvitar Singh et coll. (1973) et Kathuria (1978) ont supposé qu'il était possible d'obtenir l'information nécessaire sur les variables auxiliaires des répondants ou des unités déclarantes (UD). Cela n'est généralement pas le cas. Il peut arriver que le caractère délicat de la question ou le refus pur et simple des répondants de fournir toute information faussent l'information sur la variable auxiliaire au point de la rendre inutile. Il se peut aussi que l'information sur la variable auxiliaire ne puisse être recueillie parce que la question qui aurait permis de l'obtenir n'est pas incluse dans le questionnaire. Shvitar Singh et coll. ont également supposé que le total de population connu de la variable auxiliaire est le même à toutes les périodes. Il se peut que cela ne soit pas vrai en pratique. Si le total de population de la caractéristique principale varie d'une période à l'autre, il y a tout lieu de penser que le total de population de toute autre variable qui serait corrélée à la caractéristique principale variera également. Dans le présent document, trois estimateurs composites de total de population utilisant de l'information auxiliaire et un plan de sondage comportant des sondages successifs à deux degrés sont proposés. Les performances des trois estimateurs sont comparées empiriquement entre elles; les trois estimateurs ont également été appliqués à une enquête visant à estimer la taille moyenne des arbres dans des plantations de teck.

2. SONDAGE POUR DEUX PÉRIODES

Pour les trois estimateurs proposés, nous supposons que le total de population de la variable auxiliaire change à la deuxième période. Les estimateurs du total (ou de la moyenne) de population fondés sur un plan d'appariement partiel sont meilleurs que les estimateurs habituels du total (ou de la moyenne) de population sans appariement partiel. On peut donc s'attendre que les estimateurs proposés, T_1 , T_2 et T_3 , donneront de meilleurs résultats que l'estimateur habituel du total de population, T_0 , et que l'estimateur fondé sur le même plan d'appariement partiel mais n'utilisant pas d'information auxiliaire, T' . Dans le calcul de ces estimateurs, nous supposons que:

- (i) la taille de l'échantillon est constante à chaque période,
- (ii) la mesure normalisée P_i de la taille de la i -ième unité de sondage du premier degré (USPD) est fixée pour chaque période,
- (iii) N et M_i , les tailles de population respectives des USPD et des unités de sondage du second degré (USSD) prélevées à partir de la i -ième USPD, sont constantes dans les deux périodes,
- (iv) le total (ou la moyenne) de population de la variable auxiliaire est connu.

Les hypothèses (i) à (iii) s'appliquent à T' , T_1 , T_2 et T_3 , tandis que l'hypothèse (iv) s'applique à T_1 , T_2 et T_3 , mais non à T' et T_0 . À la première période, un échantillon S_1 de n USPD est tiré avec probabilités proportionnelles à la taille et avec remise (PPTAR) à l'aide d'une mesure normalisée de la taille de la i -ième unité ($i = 1, 2, \dots, N$). Pour choisir les USSD, nous adoptons la méthode de

Comparaison d'estimateurs de totaux de population obtenus par sondages successifs à deux degrés à l'aide de l'information auxiliaire

F.C. OKAFOR¹

RÉSUMÉ

Singh et Srivastava (1973) ont élaboré un estimateur linéaire non biaisé de moyennes de population qui pourrait être utilisé dans des sondages successifs à l'aide de plusieurs variables auxiliaires dont les moyennes de population connues ne changent pas d'une période à l'autre. Dans le présent document, trois estimateurs composites T_1 , T_2 et T_3 , utilisant chacun une variable auxiliaire dont la moyenne de population connue change d'une période à l'autre, sont présentés pour l'estimation du total de population de la période donnée. Les estimateurs sont comparés à l'estimateur habituel, T_0 , et à l'estimateur habituel de sondages successifs, T' , du total de population de la période donnée sans l'aide de l'information auxiliaire. Nous observons que l'utilisation conjuguée de l'information auxiliaire et d'une méthode par sondages successifs ne produit pas toujours uniformément un gain d'efficacité par rapport à T_0 ou T' . Toutefois, quand ils ont été appliqués à une enquête visant à estimer la taille moyenne des arbres dans des plantations de teck, les estimateurs T_1 , T_2 et T_3 se sont avérés plus efficaces que T_0 ou que T' .

MOTS CLÉS: Périodes successives; appariement partiel; variable auxiliaire.

1. INTRODUCTION

La théorie et la pratique du sondage d'une même population à des moments différents – qu'on appelle sondages échelonnés ou sondages successifs – ont été beaucoup étudiées par certains statisticiens d'enquête. Les principaux objectifs des sondages successifs sont d'estimer des paramètres de population (par exemple des totaux, des moyennes, des ratios de population, etc.) pendant la période la plus récente et d'estimer les variations dans ces paramètres d'une période à l'autre. La théorie des sondages successifs a été proposée pour la première fois par Jessen (1942). Beaucoup d'autres auteurs ont depuis étudié la question, en particulier en ce qui concerne l'estimation de moyennes de population; notamment Singh (1968), Abraham et coll. (1969), Kathuria et Singh (1971) et Kathuria (1976), pour n'en nommer que quelques-uns. Singh (1968) a été le premier à étendre la théorie des sondages à un seul degré aux sondages échelonnés à deux degrés. Il a utilisé un plan de sondage dans lequel, à la deuxième période, une fraction λ des unités de sondage du premier degré (USPD) choisies à la période précédente est retenue, en plus des unités de sondage correspondants du second degré (USSD) et d'une fraction μ ($\lambda + \mu = 1$) choisie de nouveau. Il a ensuite obtenu un estimateur non biaisé à variance minimum de la moyenne de population de la période donnée. Abraham et coll. (1969) ont considéré le cas où un appariement partiel des unités était effectué aux deux degrés. Les unités étaient choisies à l'aide d'une méthode de sondage aléatoire simple et sans remise (SASSR). Kathuria (1975) a modifié cette façon de procéder en utilisant une méthode de sélection des USPD avec probabilités proportionnelles à la taille et avec remise (PPTAR) et proposé un estimateur linéaire composite pour estimer la moyenne de population de la période donnée.

¹ F.C. Okafor, Département de statistique, Université d'Ibadan, Ibadan, Nigeria.

- McDOWELL, I. (1981). Un examen des questions proposées par l'OECD dans le cadre de l'enquête canadienne. *Revue d'épidémiologie et de santé publique*, 29, 412-429.
- MORIN, J.-P., et DOWLER, L. (1986). Proposition d'une méthodologie pour l'ESLA-institutions. Document de travail, Statistique Canada.
- MORIN, J.-P. (1986). Comparaison initiale de l'ESIC et de l'ESG. Document de travail, Statistique Canada.
- ORGANISATION MONDIALE DE LA SANTÉ, (1980). Classification internationale des déficiences, incapacités et handicaps. Genève, Suisse.
- RAJ, D. (1968). *Sampling Theory*. New York: McGraw-Hill.
- WILSON, R.W., et McNEIL, J.M. (1981). Analyse préliminaire de l'incapacité au sein de l'OCDE, d'après le test initial de l'enquête post-censitaire. *Revue d'épidémiologie et de santé publique*, 29, 469-475.

14. Éprouvez-vous des difficultés à vous habiller et vous déshabiller?
15. Éprouvez-vous des difficultés à vous mettre au lit et à sortir du lit?
16. Éprouvez-vous des difficultés à vous couper les ongles d'orteils?
17. Éprouvez-vous des difficultés à vous servir de vos doigts pour saisir ou manier un objet?
18. Éprouvez-vous des difficultés à entendre le bras dans n'importe quelle direction pour prendre quelque chose (par ex., au dessus de votre tête)?
19. Éprouvez-vous des difficultés à couper vos aliments?
20. À cause d'une affection ou un problème de santé chronique qui devrait durer 6 mois ou plus, êtes-vous limité(e) dans le genre ou la quantité d'activités que vous pouvez faire . . .
- (i) à la maison? (ii) à l'école ou au travail? (iii) dans vos autres occupations comme les déplacements, les sports ou les loisirs?
21. Un professionnel de l'enseignement ou de la santé vous a-t-il déjà dit que vous aviez des difficultés d'apprentissage?
22. De temps à autre, chacun éprouve des difficultés à se souvenir du nom d'une personne familière ou à apprendre quelque chose de nouveau ou il nous arrive d'être confus pendant quelques instants. Toutefois, avez-vous en permanence des problèmes de mémoire ou d'apprentissage?
23. À cause d'une affection ou d'un problème chronique d'ordre émotif, psychologique, nerveux ou mental, êtes-vous limité(e) dans le genre ou la quantité d'activités que vous pouvez faire . . .
- (i) à la maison? (ii) à l'école ou au travail? (iii) dans vos autres occupations comme les déplacements, les sports ou les loisirs?

BIBLIOGRAPHIE

- BEBBINGTON, A.C. (1975). A simple method of drawing a sample without replacement. *Applied Statistics*, 24, 136.
- CARTER, R.G., GILES, P.D., et SHERIDAN, M.J. (1982). Description and rationale for the screen tests for the January 1983 Disability Survey. Projet d'établissement d'une base de données sur les invalides, Division de la santé, Statistique Canada.
- CHAMBRE DES COMMUNES (1981). Obstacles, Rapport du Comité spécial concernant les invalides et les handicapés. Ottawa.
- COCHRAN, W.G. (1977). *Sampling Techniques*, (3^e éd.). New York: John Wiley.
- DOLSON, D., GILES, P., et MORIN, J.-P. (1984). Méthode d'enquête sur les personnes souffrant d'une incapacité à l'aide de questions supplémentaires de l'Enquête sur la population active. *Techniques d'enquête*, 10, 203-214.
- DOLSON, D., McCLEAN, K., MORIN, J.-P., et THÉBERGE, A. (1986). Rapport de méthodologie de l'ESLA. Document de travail, Statistique Canada.
- GRABOWIECKI, F. (1982). Discussion of the target population for the Disability Survey. Projet d'établissement d'une base de données sur les invalides, Division de la santé, Statistique Canada.
- GRABOWIECKI, F. (1983). Contenu des données de l'enquête de Statistique Canada auprès des handicapés. Rapport technique, Division de la santé, Statistique Canada.
- LAZARUS, G., et NESICH, R. (1985). A report on the methodology of the Canadian Health and Disability Survey. Document de travail, Statistique Canada.

APPENDICE

Question n° 20 du formulaire 2B du recensement

20. a) Êtes-vous limité(e) dans vos activités à cause d'une incapacité physique, d'une incapacité mentale ou d'un problème de santé chronique: (consultez le guide)

À la maison?

- ☐ non, je ne suis pas limité(e)
- ☐ oui, je suis limité(e)

À l'école ou au travail?

- ☐ non, je ne suis pas limité(e)
- ☐ oui, je suis limité(e)

☐ sans objet

Dans d'autres activités, par ex., dans vos trajets entre la maison et votre lieu de travail ou dans vos loisirs?

- ☐ non, je ne suis pas limité(e)
- ☐ oui, je suis limité(e)

b) Avez-vous des incapacités ou handicaps à long terme?

- ☐ non
- ☐ oui

Questions de sélection pour l'ESLA (questionnaire des adultes)

1. Eprouvez-vous des difficultés à entendre ce qui se dit au cours d'une conversation normale avec une autre personne?
2. Eprouvez-vous des difficultés à entendre ce qui se dit au cours d'une conversation en groupe avec au moins trois autres personnes?
4. Eprouvez-vous des difficultés à lire les caractères ordinaires d'un journal (avec des verres si vous en portez habituellement)?
5. Eprouvez-vous des difficultés à voir clairement la figure de quelqu'un à 12 pieds/4 mètres (par ex., d'un bout à l'autre d'une pièce) avec des verres si vous en portez habituellement?
7. Eprouvez-vous des difficultés à parler et être compris(e)?
8. Eprouvez-vous des difficultés à marcher sur une distance de 400 verges/mètres sans vous reposer (environ trois pâtés de maisons)?
9. Eprouvez-vous des difficultés à monter et descendre un escalier (environ 12 marches)?
10. Eprouvez-vous des difficultés à transporter un objet de 10 livres sur une distance de 30 pieds/5 kilogrammes sur 10 mètres (par ex., un sac d'épicerie)?
11. Eprouvez-vous des difficultés à vous déplacer d'une pièce à une autre?
12. Eprouvez-vous des difficultés à vous tenir debout pendant de longues périodes, c'est-à-dire pendant plus de 20 minutes? Appelez-vous qu'il s'agit de problèmes qui devraient durer 6 mois ou plus.
13. Lorsque vous êtes debout, éprouvez-vous des difficultés à vous pencher et à ramasser un objet sur le plancher (par ex., un soulier)?

6. DISCUSSION

L'enquête post-censitaire est une méthode de sondage relativement nouvelle qui est appelée à connaître de nombreux développements au cours des prochaines années. Ce type d'enquête permet beaucoup de flexibilité dans la collecte des données et l'utilisation d'échantillons de grande taille dispersés partout à travers le pays avec des coûts et des délais convenables. L'Enquête sur la santé et les limitations d'activités constitue une première expérience dans ce domaine au Canada pour une enquête de cette importance.

Le plan d'échantillonnage exposé dans le présent article représente une tentative pour maximiser l'emploi des possibilités offertes par l'approche post-censitaire avec une utilisation optimale des ressources disponibles. Le contrôle de la taille de l'échantillon demeure une des difficultés majeures inhérentes à la méthode proposée. La détermination de l'allocation de l'échantillon étant effectuée avant le recensement, tous les calculs doivent être faits à partir de projections basées sur le recensement antérieur. Dans ce contexte, la taille d'un échantillon constitué d'un ensemble de petites régions choisies suivant les résultats de ces projections risque de connaître des variations considérables une fois concrétisée dans le véritable recensement.

On se trouve alors placé devant, d'une part, la possibilité d'obtenir un échantillon insuffisant pour les exigences de qualité des estimations et, d'autre part, l'éventualité de dépasser les ressources allouées à la collecte des données. Afin d'obvier à ces difficultés, la stratégie suivante a été mise en oeuvre. Un nombre cible d'interviews pour chaque population a été calculé pour l'échantillon "oui". Ce nombre était basé sur la taille d'échantillon requise pour produire des estimés répondant à nos critères de qualité. Cependant, un nombre de SD supérieur à celui nécessaire pour obtenir le nombre cible d'interviews a été sélectionné, et cela pour les raisons mentionnées précédemment. Si le nombre réel d'interviews à faire, tel que calculé sur le terrain, est supérieur au nombre cible, alors, pour des raisons de coût, un sous-échantillon de SD est exclu de l'enquête. Ce n'est que pour le bureau régional de Halifax (recouvrant les provinces du Nouveau-Brunswick, de la Nouvelle-Écosse et de l'Île-du-Prince-Édouard) que le nombre d'interviews de l'échantillon "oui" était passablement plus élevé que le nombre cible. Il fut donc décidé d'exclure certains SD de cette partie de l'échantillon. Pour savoir quels SD seraient exclus, il fallait connaître le nombre cible et le nombre réel d'interviews pour chaque SD. Pour 40% des SD, on a dû imputer le nombre réel d'interviews car il n'était pas disponible à temps.

Pour ce faire, on connaissait le total, par district de commissaire au recensement, des nombres réels d'interviews; la portion de ce total qui n'avait pas déjà été utilisée a été distribuée parmi les SD qui requéraient une imputation, proportionnellement au nombre cible d'interviews. On a ensuite calculé la différence, pour chaque population, entre le nombre réel et le nombre cible d'interviews pour chacune des deux strates de chaque RIF. Une différence positive (réel-cible) indiquait une population pour laquelle des SD pouvaient être exclus de l'enquête. Dans chaque strate, les SD ont été classifiés en trois groupes (1, 2 et 3) selon qu'ils étaient sélectionnés pour trois, deux ou une seule des populations respectivement. Le fichier des SD a ensuite été trié par strate et par groupe de façon ascendante, l'ordre des SD à l'intérieur d'un groupe étant aléatoire. Chaque SD a été considéré successivement et était supprimé pour les trois populations si:

- 1) une différence positive demeurait non négative après suppression du SD;
- 2) une différence négative n'était pas réduite davantage.

De cette façon, chaque différence positive a été réduite à un nombre aussi près que possible de zéro compte tenu de l'ordre aléatoire des SD.

REMERCIEMENTS

Les auteurs remercient D. A. Binder pour sa contribution à l'élaboration du plan de sondage et P. Reed pour la partie informatique de l'allocation de l'échantillon. Les auteurs aimeraient également remercier les arbitres pour leurs observations utiles.

À partir des résultats d'enquêtes précédentes, on établit un lien entre la pyramide des âges de la population d'un SD et le nombre de personnes handicapées qu'on peut s'attendre à trouver dans ce SD. Le nombre de personnes handicapées étant inconnu, la variable utilisée pour la stratification et l'allocation de l'échantillon est le nombre attendu de personnes handicapées.

Dans le cas présent, on n'a que deux strates: celle des grands SD et celle des petits SD. Puisqu'on utilise l'allocation proportionnelle, pour déterminer d'une manière optimale la taille frontière au-delà de laquelle un SD est grand et en deçà de laquelle il est petit, on s'est servi d'un critère qu'on peut trouver dans Raj (1968). La taille frontière doit être égale à la moyenne de la taille moyenne des petits SD et de la taille moyenne des grands SD.

5. SÉLECTION DE L'ÉCHANTILLON

Il fallait choisir des échantillons pour les trois populations (enfants, adultes et personnes âgées) parmi les grands et les petits SD "oui" de chaque RIP, tant pour la base trois que pour la base cinq, et parmi les SD "non" de chaque province pour la base cinq. Lorsque, dans une RIP, il y avait moins de deux grands SD ou moins de deux petits SD, on sélectionnait tous les SD de cette RIP pour les trois populations. Les échantillons "oui" et "non" étaient créés indépendamment en utilisant l'algorithme à un passage décrit par Bebbington (1975). Les échantillons des trois populations, pour la composante "oui" et pour la composante "non", étaient emboîtés afin de minimiser le nombre total de SD choisis.

Au tableau suivant figurent les tailles des échantillons par province pour chaque groupe d'âge.

Tableau 2
Tailles d'échantillons par province et par groupe d'âge

Province	Enfants			Adultes			Personnes âgées		
	Nombre de SD "oui"	Nombre de SD "non"	sélectionnés	Nombre de SD "oui"	Nombre de SD "non"	sélectionnés	Nombre de SD "oui"	Nombre de SD "non"	sélectionnés
T.-N.	880	136	405	154	476	173			
I.P.-E.	242	242	111	217	82	166			
N.-E.	1257	157	434	130	438	115			
N.B.	1142	162	459	146	453	138			
Qué.	4749	153	1070	114	1488	133			
Ont.	6085	158	1304	116	1542	120			
Man.	1082	203	457	169	367	144			
Sask.	2291	265	942	241	921	193			
Alb.	2762	190	909	176	1389	222			
C.-B.	3117	170	752	125	948	119			

Tableau 1

Comparaison des quatre modèles

Prov.	EFF_S	EFF_M	EFF_A
T.-N.	0.890	0.891	0.887
I.P.-E.	0.903	0.914	0.919
N.-E.	0.960	0.972	0.912
N.-B.	0.869	0.868	0.884
Qué.	0.778	0.764	0.818
Ont.	0.932	0.930	0.916
Man.	0.892	0.904	0.912
Sask.	0.732	0.749	0.801
Alb.	0.818	0.827	0.860
C.-B.	0.713	0.716	0.775
Yukon	0.770	0.768	0.840
T.N.-O.	1.252	1.246	1.157

On a ensuite calculé pour chaque province p , une mesure d'efficacité des modèles simple, multiplicatif et additif par rapport au modèle trivial:

$$EFF_m = \frac{\sum_{k=1}^{N_p} \sum_{l=1}^{N_k} \sum_{j=1}^{16} \left(\hat{EC}_{l,k;81}^{(m)}(j) - EC_{l,k;81}(j) \right)^2}{\sum_{k=1}^{N_p} \sum_{l=1}^{N_k} \sum_{j=1}^{16} \left(\hat{EC}_{l,k;81}^{(T)}(j) - EC_{l,k;81}(j) \right)^2} \quad (m = S, M, A).$$

où $\hat{EC}_{l,k;81}^{(m)}(j)$ avec $m = S, M, A$ et T sont les projections obtenues à l'aide des modèles simple, multiplicatif, additif et trivial respectivement. Le tableau 1 présente des valeurs obtenues.

Le modèle simple donne les pires résultats pour une province et une territoire, le modèle multiplicatif pour deux provinces, et le modèle additif pour sept provinces et un territoire. Le modèle simple se révèle le meilleur pour cinq provinces, le modèle multiplicatif pour deux provinces et un territoire, et le modèle additif pour trois provinces et un territoire.

Le modèle simple ayant aussi, comme son nom l'indique, l'avantage d'être simple, c'est le modèle qui a été retenu.

4.4 Méthode de stratification selon la taille des secteurs de dénombrement

Si on faisait un tirage aléatoire simple pour sélectionner les SD à l'intérieur de chaque région infra-provinciale (RIP), alors les personnes handicapées faisant partie d'un SD avec de nombreuses personnes handicapées auraient moins de chance d'être sélectionnées que celles faisant partie d'un petit SD, c'est-à-dire un SD avec peu de personnes handicapées. Pour éviter de trop grandes différences au niveau des probabilités de sélection, dans chaque RIP on stratifie la population des SD selon le nombre de personnes handicapées qu'il y trouve, puis on utilise l'allocation proportionnelle, c'est-à-dire que dans chaque strate, le nombre de SD choisis est proportionnel au nombre de personnes handicapées.

Le troisième modèle utilisé, appelé modèle additif, est donné par les équations suivantes:

$$\begin{aligned} \hat{EC}_{l,k;86}(j) &= EC_{l,k;81}(j) + e_l + f_j \quad (l = 1, \dots, N_k; j = 1, \dots, 16), \\ \sum_{N_k}^{l=1} \hat{EC}_{l,k;86}(j) &= \hat{DR}_{k;86}(j) + D \quad (j = 1, \dots, 16), \end{aligned}$$

$$\sum_{N_k}^{l=1} e_l = D,$$

$$\sum_{16}^{j=1} \hat{EC}_{l,k;86}(j) = \hat{EC}_{l,k;86}(tot) \quad (l = 1, \dots, N_k).$$

On suppose donc que les accroissements (ou diminutions) de population pour chaque groupe d'âge des ensembles correspondants de la DR peuvent se décomposer en deux termes: un qui ne dépend que de l'ensemble correspondant et non de l'âge (e_l), et un qui ne dépend que de l'âge et pas de l'ensemble correspondant (f_j).

Un dernier modèle, trivial celui-là, serait simplement de poser

$$\hat{EC}_{l,k;86}(j) = EC_{l,k;81}(j) \quad (l = 1, 2, \dots, N_k; j = 1, \dots, 16).$$

4.3 Évaluation des méthodes d'estimation

Les quatre méthodes ont été évaluées en utilisant des données pour la période allant de 1976 à 1981. On a utilisé la projection établie en 1976 de la population par groupe d'âge et par province en 1981 ($P_{81}(j)$), la population par groupe d'âge et par SD en 1976, un fichier de conversion 1976-1981 et l'estimé pré-censitaire du nombre de logements par SD en 1981. Des estimations pour la population par groupe d'âge et par DR en 1979 (l'équivalent de $DR_{k;84}(j)$) n'existant pas, on a posé

$$\hat{DR}_{k;81} = \frac{P_{81}(j) \hat{DR}_{k;84}(j)}{\sum_{N_p}^{k=1} \hat{DR}_{k;84}(j)}.$$

Pour $\hat{EC}_{l,k;81}(tot)$, nécessaire pour les modèles multiplicatif et additif, on a utilisé

$$\hat{EC}_{l,k;81}(tot) = \frac{\sum_{16}^{j=1} EC_{l,k;76}(j)}{\sum_{N_k}^{j=1} EC_{l,k;76}(j)}.$$

La première méthode pour estimer $EC_{l,k;86}(j)$ consiste à supposer l'existence de K_j ($j = 1, \dots, 16$) tels que

$$\begin{aligned} \hat{EC}_{l,k;86}(j) &= K_j EC_{l,k;81}(j) \quad (l = 1, \dots, N_k; j = 1, \dots, 16), \\ \sum_{N_k}^{l=1} \hat{EC}_{l,k;86}(j) &= \hat{DR}_{k;86}(j) \quad (j = 1, 2, \dots, 16). \end{aligned}$$

On dira que cette méthode utilise le modèle simple. On obtiendra

$$\hat{EC}_{l,k;86}(j) = \frac{\sum_{N_k}^{l=1} EC_{l,k;81}(j)}{\hat{DR}_{k;86}(j) EC_{l,k;81}(j)} \quad (l = 1, \dots, N_k; j = 1, \dots, 16).$$

Avec ce modèle simple, l'estimation de la population totale de $EC_{l,k}$ en 1986 est

$$\sum_{j=1}^{16} \frac{\sum_{N_k}^{l=1} EC_{l,k;81}(j)}{\hat{DR}_{k;86}(j) EC_{l,k;81}(j)}.$$

Si on croit être en mesure de fournir une meilleure estimation, $\hat{EC}_{l,k;86}(tot)$ de cette quantité par des moyens indépendants (par exemple, à partir de l'estimé du nombre de logements dans $EC_{l,k}$ en 1986), alors on peut utiliser des modèles plus élaborés pour estimer $EC_{l,k;86}(j)$. Le modèle multiplicatif est précisé par les équations suivantes:

$$\begin{aligned} \hat{EC}_{l,k;86}(j) &= K_j (EC_{l,k;81}(j)) + e'_l \quad (l = 1, \dots, N_k; j = 1, \dots, 16), \\ \sum_{N_k}^{l=1} \hat{EC}_{l,k;86}(j) &= K (DR_{k;86}(j)) \quad (j = 1, \dots, 16), \\ \sum_{N_k}^{l=1} e'_l &= 0, \end{aligned}$$

$$\sum_{j=1}^{16} \hat{EC}_{l,k;86}(j) = \hat{EC}_{l,k;86}(tot) \quad (l = 1, \dots, N_k).$$

On peut interpréter e_l comme étant la migration intra-DR nette du l -ième ensemble correspondant.

4. FICHIER GÉOGRAPHIQUE ET DÉMOGRAPHIQUE POUR 1986

4.1 Description des informations disponibles

Lors de l'allocation de l'échantillon qui a eu lieu au printemps de 1986, on disposait des données suivantes pour établir les projections de la population par groupe d'âge et par SD:

1. projection de la population par groupe d'âge et par province en 1986;
2. estimé de la population par groupe d'âge et par DR en 1984;
3. population par groupe d'âge et par SD en 1981;
4. un fichier de conversion pour établir la correspondance entre les SD de 1981 et ceux de 1986;
5. estimé du nombre de logements par SD en 1986.

La façon dont est construit le fichier de conversion repose sur le concept d'ensembles correspondants. Chaque ensemble correspondant est la plus petite région formée de SD, dont les frontières sont demeurées inchangées. Par exemple, si trois SD de 1981 ont servi à former deux SD de 1986, ce groupe de trois SD de 1981 (ou deux SD de 1986) forme un ensemble correspondant. Les quatre méthodes décrites dans la sous-section qui suit visent à établir des projections de la population par groupe d'âge et par ensemble correspondant en 1986. Si un ensemble correspondant est formé de plusieurs SD de 1986, la population projetée pour cet ensemble correspond au nombre de logements par SD en 1986.

4.2 Méthodes d'estimation

Pour la province p , notons

EC_{lk} = le l -ième ensemble correspondant de la k -ième DR ($l = 1, 2, \dots, N_k$;

$k = 1, 2, \dots, N_p$),

$EC_{l,k;81}(j)$ = population de EC_{lk} dans le j -ième groupe d'âge en 1981 ($j = 1, 2, \dots, 16$),
 $DR_{k;84}(j)$ = estimé de la population de la k -ième DR dans le j -ième groupe d'âge en 1984,
 $P_{86}(j)$ = projection de la population dans le j -ième groupe d'âge dans la province en 1986.

Pour les trois méthodes qui suivent, la première étape consiste à calculer $\hat{DR}_{k;86}(j)$, c'est-à-dire la projection de la population de la k -ième DR dans le j -ième groupe d'âge en 1986. On suppose qu'il existe des K_j ($j = 1, 2, \dots, 16$) tels que

$$\hat{DR}_{k;86}(j) = K_j(DR_{k;84}(j)) \quad (k = 1, 2, \dots, N_p; j = 1, 2, \dots, 16),$$

$$\sum_{p=1}^{N_p} \hat{DR}_{k;86}(j) = P_{86}(j) \quad (j = 1, 2, \dots, 16).$$

Cela implique que

$$\hat{DR}_{k;86} = \frac{\sum_{p=1}^{N_p} DR_{k;84}(j)}{P_{86}(j) DR_{k;84}(j)}.$$

Quels sont les ensembles E_1 , E_2 , E_3 et E_4 qui correspondent à la solution? L'ensemble E_4 est facile à déterminer. Par ailleurs, on doit avoir

$$a_k < (d_k/c_k)^{1/2} K < b_k \quad (k \in E_3), \quad (d_k/c_k)^{1/2} K \geq b_k \quad (k \in E_2),$$

$$(3.9) \quad (d_k/c_k)^{1/2} K \leq a_k \quad (k \in E_1).$$

Il s'agirait donc d'essayer chacune des possibilités pour E_1 , E_2 et E_3 jusqu'à ce qu'on obtienne une valeur de K qui satisfasse (3.9). Pour réduire le nombre de possibilités à examiner, remarquons que si pour $k' \geq k$,

$$(3.10) \quad b_k (c_k/d_k)^{1/2} \geq b_{k'} (c_{k'}/d_{k'})^{1/2} \quad (k, k' \in \{0, 1, \dots, N_p\}),$$

alors il existe k^* tel que $E_2 = \{0, 1, 2, \dots, k^*\}$, ou bien $E_2 = \{ \}$. Tandis que si pour $k' \geq k$,

$$(3.11) \quad a_k (c_k/d_k)^{1/2} \geq a_{k'} (c_{k'}/d_{k'})^{1/2} \quad (k, k' \in \{0, 1, \dots, N_p\}),$$

alors il existe k^{**} tel que $E_1 = \{k^{**}, k^{**} + 1, \dots, N_p\}$ ou bien $E_1 = \{ \}$.

3.3 Estimation des paramètres

Afin d'effectuer le calcul de l'allocation optimale de l'échantillon, les quantités suivantes doivent être évaluées:

P_1 = proportion des individus ayant été sélectionnés lors de l'ESLA et qui ont répondu "oui" à la question 20 du recensement;

P_2 = proportion des individus n'ayant pas été sélectionnés lors de l'ESLA et qui ont répondu "oui" à la question 20 du recensement;

P_3 = proportion des individus ayant été sélectionnés lors de l'ESLA et qui ont répondu "non" à la question 20 du recensement.

Puisque ces paramètres ne peuvent être calculés directement à partir des données de l'Enquête sur la santé et les incapacités au Canada, un test qu'on a appelé "l'étude de calibration" a été effectué en septembre et octobre 1985.

La question 20 du recensement fut posée intégralement comme question supplémentaire lors de l'Enquête sur la population active (EPA) de septembre. Elle s'adressait à un échantillon d'environ 36,000 individus. Les questions sur les 17 AVQ et une question sur les handicaps mentaux constituaient un supplément à l'EPA d'octobre et elles étaient demandées aux mêmes individus.

Pour chacun des groupes d'âges de cinq ans, on s'est servi des valeurs pondérées de l'étude de calibration afin d'estimer la probabilité d'une réponse positive P (oui) à la question 20 du recensement. Le questionnaire de sélection de l'ESLA diffère de celui de l'étude de calibration. Dans l'ESLA, il y a plus de questions sur les troubles mentaux ou psychologiques et la partie (a) de la question 20 du recensement est posée de nouveau. Donc, on ne s'est pas appuyé seulement sur l'étude de calibration pour calculer les paramètres.

tandis que si $\lambda_k n_k > N_{2k}$ alors on prend $n_k = N_{2k}$ et

$$n_k = n_k + \frac{N_{2k}}{(\lambda_k n_k - N_{2k}) N_{1k} \lambda_k}.$$

On considère donc que $N_{2k} / (N_{1k} \lambda_k)$ petits SD équivalent à un grand SD. Il y a, en moyenne, autant de personnes handicapées dans un grand SD que dans $N_{2k} / (N_{1k} \lambda_k)$ petits. En procédant ainsi, on ne respecte pas toujours la relation $n_{2k} = \lambda_k n_{1k}$. Toutefois, on évite de se contenter d'un CV supérieur à celui visé, lorsque, par exemple, il reste des petits SD à observer (même si tous les grands SD ont été sélectionnés).
Il se peut que pour certaines valeurs de k , on ait $a_k \geq b_k$. Dans un tel cas, on prend $n_k = b_k$. Soient

$$\begin{aligned} E_1 &= \{k = 0, 1, 2, \dots, N_p \mid n_k = a_k\}, \\ E_2 &= \{k = 0, 1, 2, \dots, N_p \mid n_k = b_k > a_k\}, \\ E_3 &= \{k = 0, 1, 2, \dots, N_p \mid a_k < n_k < b_k\}, \\ E_4 &= \{k = 0, 1, 2, \dots, N_p \mid n_k = b_k \leq a_k\}, \end{aligned}$$

la solution existe si

$$\sum_{N_p}^{k=0} d_k / b_k \leq e,$$

et elle est de la forme

$$n_k = \begin{cases} a_k & (k \in E_1) \\ b_k & (k \in E_2 \cup E_4) \\ K(d_k / c_k)^{1/2} & (k \in E_3) \end{cases} \quad (3.7)$$

où

$$K = \frac{\sum_{k \in E_3} (d_k / c_k)^{1/2}}{\sum_{k \in E_1} d_k / a_k - \sum_{k \in E_2 \cup E_4} d_k / b_k}, \quad (3.8)$$

puisque les n_k ($k \in E_3$) minimisent $\sum_{k \in E_3} c_k n_k$, sous la contrainte

$$\sum_{k \in E_1} d_k / a_k - \sum_{k \in E_2 \cup E_4} d_k / b_k = e.$$

En considérant l'équation (3.1), on a

$$(3.4) \qquad \text{Var}(t_1) = \sum_{N_p}^{k=1} \left(\frac{A_k Y_k}{n_{1k}} - B_k Y_k \right).$$

De (3.2), (3.3) et (3.4), on obtient

$$\text{Var}(B) = B_2 \left\{ \left(\frac{A_0}{n_0 T_0^2} - \frac{B_0}{T_0^2} \right) + \sum_{N_p}^{k=1} \left(\frac{A_k Y_k}{n_{1k} T_1^2} - \frac{B_k Y_k}{T_1^2} \right) \right\}$$

$$= \frac{1}{n_0} \left(\frac{B_2 A_0}{T_0^2} \right) + \sum_{N_p}^{k=1} \frac{1}{n_{1k}} \left(\frac{B_2 A_k Y_k}{T_1^2} \right) - B_2 \left(\frac{T_0^2}{B_0} + \sum_{N_p}^{k=1} \frac{B_k Y_k}{T_1^2} \right)$$

$$= \frac{X}{n_0} + \sum_{N_p}^{k=1} \frac{n_k}{W_k} - Z.$$

On peut réexprimer le problème d'optimisation comme étant celui de minimiser

$$\sum_{N_p}^{k=0} c_k n_k$$

en tenant compte des contraintes

$$(3.5) \qquad 0 < a_k \leq n_k \leq b_k \quad (k = 0, 1, 2, \dots, N_p)$$

et

$$(3.6) \qquad \sum_{N_p}^{k=0} d_k / n_k = e$$

où, pour $k = 1, 2, \dots, N_p$,

$$n_k = n_{1k}, \quad c_k = c_{1k} + c_{2k} \lambda_k, \quad a_k = \frac{A_k CV_k^* + B_k}{A_k}, \quad b_k = \min(N_{1k}, N_{2k} / \lambda_k).$$

En pratique, plutôt que de prendre $b_k = \min(N_{1k}, N_{2k} / \lambda_k)$ on prend

$$b_k = \frac{N_{1k} N_{2k} (1 + \lambda_k)}{\lambda_k^2 N_{1k} + N_{2k}},$$

puis si $n_k > N_{1k}$, alors on prend $n_{1k} = N_{1k}$ et

$$n_{2k} = \lambda_k n_k + \frac{N_{1k} \lambda_k}{(n_k - N_{1k}) N_{2k}},$$

On peut donc représenter $CV^2(y_k)$ sous la forme suivante:

$$CV^2(y_k) = \frac{Y_k^2}{A_k} = \frac{n_{1k}}{A_k} - B_k.$$

(3.1)

Par ailleurs, B , le biais relatif, et B , son estimateur, sont donnés par

$$B = \frac{T_0}{T_1} = \frac{\sum_{i=1}^{N_0} M_{i0} p_{i0}}{\sum_{k=1}^{N_d} Y_k},$$

$$B = \frac{t_1}{t_0} = \frac{\sum_{i=1}^{n_0} \frac{N_0}{M_{i0} p_{i0}} u_{i0}}{\sum_{k=1}^{N_d} y_k},$$

où M_{i0} est le nombre de personnes "non" dans le SD i de la province, et p_{i0} est la probabilité d'une caractéristique d'intérêt pour une personne "non" du SD i .

En supposant que t_0 et t_1 sont indépendants, on a

$$\text{Var}(B) = B^2 \left(\frac{T_0^2}{\text{Var}(t_0)} + \frac{T_1^2}{\text{Var}(t_1)} \right).$$

(3.2)

Après quelques opérations algébriques, si f_0 est la fraction de sondage dans les SD "non", on obtient

$$\text{Var}(t_0) = \frac{1}{N_0} \left\{ N_0^2 S_2 + N_0 \left(\frac{f_0}{1 - f_0} \right) \left(\sum_{i=1}^{N_0} M_{i0} S_i^2 \right) - N_0 S_2 \right\}$$

où

$$S_2 = \frac{1}{N_0} \sum_{i=1}^{N_0} M_{i0} p_{i0} - \left(\sum_{i=1}^{N_0} \frac{M_{i0}^2}{M_{i0} p_{i0}} \right)^2, S_2^i = \frac{M_{i0}^2}{M_{i0} p_{i0}} (1 - p_{i0})$$

ce qui peut être représenté sous la forme suivante:

$$\text{Var}(t_0) = \frac{n_0}{A_0} - B_0.$$

(3.3)

D'autre part, en supposant les y_k indépendants, on a

$$\text{Var}(t_1) = \sum_{k=1}^{N_d} \text{Var}(y_k).$$

Notons N_0 le nombre de SD "non" dans la province, N_{jk} le nombre de SD "oui" dans la strate j et la RIP k de la province, n_0 et n_{jk} les tailles d'échantillons correspondantes et c_0 et c_{jk} les coûts unitaires d'échantillonnage correspondants. Si on a N_p RIP dans la province, on veut donc minimiser

$$\sum_{k=1}^{N_p} (c_{1k} n_{1k} + c_{2k} n_{2k}) + c_0 n_0$$

étant donné

$$CV^2(y_k) \leq CV^2_*; \text{Var}(B) = \text{Var}_*(B);$$

$$n_{jk} \leq N_{jk}; n_{2k} = \lambda_k n_{1k}; n_0 \leq N_0$$

$$(j = 1, 2; k = 1, \dots, N_p)$$

où λ_k est le rapport du nombre attendu de personnes handicapées dans les petits SD sur le nombre attendu de personnes handicapées dans les grand SD de la RIP k , y_k est le nombre estimé de personnes "oui" qui ont une caractéristique d'intérêt dans la RIP k et les valeurs notées avec des * sont des constantes.

Si la fraction de sondage dans les SD "oui" est f_j , M_{ijk} est le nombre de personnes "oui" dans le SD i de la strate j de la RIP k de la province, et p_{ijk} est la probabilité d'une caractéristique d'intérêt pour une personne "oui" du SD i de la strate j de la RIP k , alors on a

$$E(y_k) = Y_k = \sum_{j=1}^2 \sum_{k=1}^{N_{jk}} M_{ijk} p_{ijk},$$

$$\text{Var}(y_k) = \sum_{j=1}^2 \left\{ \frac{n_{jk}^2}{N_{jk}^2} \left(1 - \frac{n_{jk}}{N_{jk}} \right) S_{2jk}^2 + \frac{n_{jk}}{N_{jk}} \left(1 - \frac{f_1}{f_1} \right) \sum_{i=1}^{N_{jk}} M_{ijk} S_{2jk}^2 \right\},$$

où

$$y_k = \sum_{j=1}^2 \sum_{k=1}^{N_{jk}} \frac{n_{jk}}{N_{jk}} M_{ijk} p_{ijk},$$

$$S_{2jk}^2 = \frac{1}{N_{jk}^2} \left(\sum_{i=1}^{N_{jk}} M_{ijk} p_{ijk} - \left(\sum_{i=1}^{N_{jk}} \frac{M_{ijk} p_{ijk}}{N_{jk}} \right)^2 \right),$$

$$S_{2jk}^2 = \frac{M_{ijk}^2}{M_{ijk}^2} \frac{1}{p_{ijk}} \left(1 - p_{ijk} \right).$$

Après quelques opérations algébriques, on obtient

$$\text{Var}(y_k) = \frac{1}{N_{jk}^2} S_{1k}^2 + \left(1 - \frac{f_1}{f_1} \right) N_{1k} \sum_{i=1}^{N_{1k}} M_{i1k} S_{2i1k}^2 + \frac{\lambda_k}{N_{2k}^2 S_{2k}^2} + \left(1 - \frac{f_1}{f_1} \right) \frac{\lambda_k}{N_{2k}^2} \sum_{j=1}^2 N_{jk} S_{2jk}^2.$$

3. PLAN DE SONDAGE

La méthode d'échantillonnage dont il est question dans cette section a été utilisée pour les bases de sondage trois et cinq. Pour des raisons d'espace, le plan de sondage relatif à la deuxième base de sondage n'est pas exposé dans le présent article. (Pour plus de renseignements sur la méthodologie de l'ESLA, voir Dolson et coll. (1986)).

3.1 Plan d'échantillonnage

Chaque province est partitionnée en régions infra-provinciales (RIP) lesquelles sont elles-mêmes partitionnées en secteurs de dénombrement (SD).

Les habitants de chaque SD sont divisés en un SD "oui" et un SD "non", selon qu'ils répondraient positivement ou négativement à la question 20 du formulaire du recensement. Dans chaque RIP, on stratifie les SD "oui" en grands et petits SD suivant le critère expliqué dans la quatrième section du présent article. Les personnes appartenant à un SD "oui" sont associées à une strate et à une RIP alors que les personnes appartenant à un SD "non" ne sont associées qu'à leur SD. Dans chaque province, la population est subdivisée en trois groupes d'âges: enfants (moins de 15 ans), adultes (15 à 64 ans) et personnes âgées (65 ans et plus). La méthode d'échantillonnage est un plan de sondage stratifié à deux degrés pour les SD "oui" de chaque RIP et un plan de sondage à deux degrés pour les SD "non" de la province. Les SD constituaient les unités primaires et les répondants les unités secondaires. Toutes les personnes ayant répondu au formulaire 2B du recensement dans un SD "oui" sélectionné dans l'échantillon sont interviewées et un tiers d'entre elles le sont dans les SD "non" sélectionnés dans l'échantillon.

3.2 Allocation de l'échantillon

Ce plan de sondage doit permettre de minimiser les coûts d'échantillonnage étant donné un coefficient de variation maximum des estimations et une variance fixée de l'estimateur B du biais relatif B . On définit B comme étant le rapport du nombre T_0 de personnes "non" qui ont une caractéristique d'intérêt dans la province, sur le nombre T_1 de personnes "oui" qui ont une caractéristique d'intérêt dans la province. Par personne "non" il faut entendre une personne qui répondrait négativement à tous les volets de la question 20 du formulaire du recensement, et par personne "oui", une personne qui répondrait positivement à au moins un des volets.

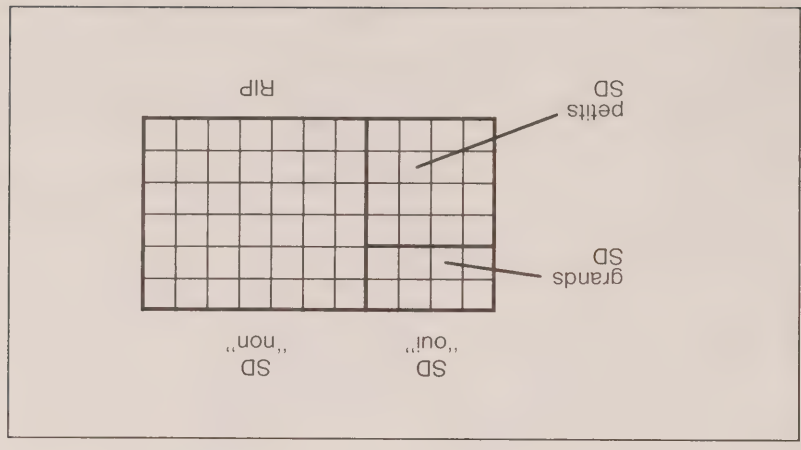


Figure 1. Illustration du plan de sondage.

La seconde étape est l'ESLA. On procède à des interviews individuelles pour la strate des "oui" et à des interviews téléphoniques pour la strate des "non". Du point de vue opérationnel, l'interview est formée de deux parties: le questionnaire de sélection et le suivi.

Le rôle du questionnaire de sélection est de déterminer pour quel répondant le questionnaire de suivi est pertinent. Le questionnaire destiné aux adultes comprend les 17 activités de la vie quotidiennes (AVQ) utilisées lors de l'Enquête sur la santé et les incapacités au Canada qui a eu lieu en 1983 et 1984 ainsi que la partie (a) de la question 20 du recensement et quelques questions sur les maladies et les handicaps mentaux (voir l'appendice B). Si une réponse positive est obtenue à au moins une de ces questions, l'interviewer procède au suivi, sinon l'interview prend fin. La partie (a) de la question du recensement est posée de nouveau pour tenir compte d'un éventuel changement de situation causé soit par une réponse d'un intermédiaire lors du recensement, soit par une réévaluation du répondant de son propre état.

La section de sélection sur le questionnaire destinée aux enfants comprend des questions sur les appareils spéciaux, les limitations aux activités, la fréquentation d'une école spéciale et les affections ou problèmes de santé. Une réponse positive à au moins une de ces questions entraîne l'interview de suivi. La question du recensement n'est pas posée de nouveau parce que tous les interviews relatives aux enfants nécessitent un intermédiaire et que la question sur les limitations d'activités est équivalente à la partie (a) de la question 20 du recensement.

La deuxième section du présent article décrit comment la population canadienne a été subdivisée en différentes sous-populations pour fins d'estimation. La section trois est consacrée au plan de sondage de l'ESLA. La section suivante traite du fichier des données géographiques et démographiques projetées pour 1986 qui a servi à former les bases de sondage. Enfin la dernière section décrit comment on a procédé à l'échantillonnage.

2. POPULATIONS COUVERTES

Les personnes qui résident en permanence dans les hôpitaux généraux, les hôpitaux psychiatriques, les centres de soins spéciaux ou les établissements pour personnes âgées et malades chroniques, les centres de traitement ou établissements pour handicapés physiques et les orphelinats ou foyers pour enfants font l'objet d'une partie spéciale de l'enquête: l'ESLA-institutions. Dans le présent article, il est question de la partie de l'enquête qui couvre la population canadienne à l'exclusion des personnes couvertes par l'ESLA-institutions, des personnes qui résident dans les prisons, les camps militaires, les foyers pour les jeunes délinquants, les navires de guerre, les établissements pénitentiaires, les établissements de correction et des résidents des logements collectifs de la catégorie "autres" (par exemple, cirques, communes non-religieuses). Chaque secteur de dénombrement (SD) dont la population n'est pas totalement exclue de l'enquête est classé dans une des cinq bases de sondage suivantes:

1. Réserves indiennes où le dénombrement avait été fait au moyen d'interviews en 1981;
2. Autres réserves indiennes;
3. SD où le dénombrement se fait habituellement par interviews;
4. SD dans les RIP de Whitehorse, Yellowknife, Pine Point, Hay River et Fort Smith;
5. Tous les autres SD.

L'ordre de priorité pour l'appartenance à une base est 1-2-4-3-5. Ainsi, un SD formé par une réserve indienne situé dans la RIP de Whitehorse est classé comme étant une réserve indienne. Chaque SD est divisé en deux: le SD "oui", formé des personnes qui donneraient une réponse positive à la question du recensement, et le SD "non" formé des personnes qui y répondraient négativement. Un plan d'échantillonnage différent est utilisé pour chacune des cinq bases de sondage: dans la première base, tous les SD "oui" sont sélectionnés et aucun des SD "non"; dans la base deux, tous les SD "oui" et un échantillon des SD "non" sont sélectionnés; dans la troisième base, aucun des SD "non" n'est choisi et un échantillon des SD "oui" sont sélectionnés; tous les SD de la base quatre sont sélectionnés; et enfin un échantillon des SD "oui" et un échantillon des SD "non" sont choisis dans la base cinq.

Plan d'échantillonnage pour l'enquête sur la santé et les limitations d'activités

D. DOLSON, K. McCLEAN, J.-P. MORIN, et A. THÉBERGE¹

RÉSUMÉ

L'Enquête sur la santé et les limitations d'activités s'inscrit dans le cadre du programme visant à établir une base de données relative à la population ayant une incapacité ou un handicap au Canada. On décrit le plan d'échantillonnage utilisé pour la partie de l'enquête couvrant la population vivant hors des institutions. Les méthodes employées pour déterminer les tailles des échantillons et pour la sélection de ceux-ci sont aussi exposées.

MOTS CLÉS: Incapacité; échantillonnage stratifié; échantillonnage à deux degrés; allocation optimale; échantillonnage sans remise.

1. INTRODUCTION

Dans le cadre du programme visant à obtenir plus d'information sur la population des personnes handicapées au Canada, l'Enquête sur la santé et les limitations d'activités (ESLA) a été réalisée au cours de l'automne 1986. Elle vise à recueillir des informations sur la nature des problèmes éprouvés par cette population et, d'une façon générale, sur leurs activités quotidiennes (à la maison, au travail, à l'école, dans les déplacements, etc.). L'enquête se divise en deux parties: l'une couvre la population vivant en institution et l'autre, qui fait l'objet du présent article, couvre la population hors des institutions.

On a partitionné le Canada en 238 régions infra-provinciales (RIP). Ces régions comprennent toutes les municipalités du Québec et de l'Ontario qui regroupent plus de 125,000 habitants et toutes celles des autres provinces de plus de 75,000 habitants. Les autres régions sont composées d'agglomérations de sous-divisions de recensement respectant la contiguïté géographique et les frontières entre les provinces. Le nombre de ces régions dans chaque province est proportionnel à la racine carrée de la population moins les municipalités précédemment définies. Un des principaux objectifs de l'enquête est de produire au niveau des RIP des statistiques sur la population handicapée afin de permettre une analyse détaillée des divers besoins. En outre, on fournira des estimations pour trois groupes d'âge: enfants (moins de 15 ans), adultes (de 15 à 64 ans) et personnes âgées (65 ans et plus).

La cueillette des informations s'est effectuée en deux étapes. La première consiste en une question à plusieurs volets, incluse dans le formulaire 2B du recensement de la population du Canada de 1986, qui porte sur les limitations du répondant dans divers secteurs d'activités ainsi que sur sa propre évaluation de sa condition. Une copie de cette question (question n° 20 du formulaire du recensement de 1986) est donnée en appendice. La seconde étape a eu lieu quelque temps après le recensement. Elle est constituée d'un questionnaire de sélection et d'un suivi qui recueille des informations sur les problèmes et les activités des répondants handicapés. La première étape vise essentiellement à répartir les répondants en deux groupes: ceux qui ont répondu "oui" à au moins une des parties de la question 20 et ceux qui ont répondu "non" à toutes les parties. Le but est d'identifier à l'avance une grande partie de la population potentiellement handicapée afin de concentrer le maximum des ressources de l'enquête sur le groupe cible. Notons toutefois que les enquêtes précédentes montrent que le groupe cible ne sera pas complètement identifié par cette question. (Voir Dolson et coll. (1984) et Dolson et coll. (1986)).

¹ D. Dolson, K. McClean, J.-P. Morin, et A. Théberge, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa, Ontario, K1A 0T6.

toute évidence, quand une des $(n + 1)$ unités de l'échantillon est rejetée au hasard, l'échantillon définitif comprend deux unités de chacun de $(n - 1) / 2$ groupes et seulement une unité d'un des groupes. Un estimateur non biaisé et positif de $Var(X)$, semblable à celui de l'expression (3.1), peut ensuite être calculé à partir des $(n - 1) / 2$ groupes qui contiennent deux unités de l'échantillon.

REMERCIEMENTS

Les auteurs remercient l'arbitre pour ses conseils pertinents concernant la première version.

BIBLIOGRAPHIE

BREWER, K.R.W. (1963). A model of systematic sampling with unequal probabilities. *Australian Journal of Statistics*, 5, 5-13.

BREWER, K.R.W. et HANIF, M. (1983). *Sampling with Unequal Probabilities*. Lecture Notes in Statistics, No. 15, New York: Springer-Verlag.

COCHRAN, W.G. (1963). *Sampling Techniques*, (2^e éd.). New York: John Wiley.

DES RAJ (1965). Variance estimation in randomized systematic sampling with probability proportional to size, *Journal of the American Statistical Association*, 60, 278-284.

DES RAJ (1966). Some remarks on a simple procedure of sampling without replacement. *Journal of the American Statistical Association*, 61, 391-396.

DURBIN, J. (1967). Design of multi-stage surveys for the estimation of sampling errors. *Journal of the Royal Statistical Society, Sér. C*, 16, 152-164.

HANURAV, T. (1967). Optimum utilization of auxiliary information: pps sampling of two units from a stratum. *Journal of the Royal Statistical Society, Sér. B*, 29, 379-391.

HORVITZ, D.G. et THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

MIDZUNO, H. (1951). On the sampling system with probability proportionate to sum of sizes. *Annals of the Institute of Statistical Mathematics*, 2, 99-108.

RAO, J.N.K. (1963). On three procedures of unequal probability sampling without replacement. *Journal of the American Statistical Association*, 58, 202-215.

RAO, J.N.K. et BAYLESS, D.L. (1969). An empirical study of the stabilities of estimators and variance estimators in unequal probability sampling of two units by stratum. *Journal of the American Statistical Association*, 64, 540-559.

RAO, J.N.K. et LANKE, J. (1984). Simplified unbiased variance estimation for multistage designs. *Biometrika*, 71, 387-395.

SAMPFORD, M.R. (1962). *An Introduction to Sampling Theory*. Edimbourg: Oliver and Boyd.

SAMPFORD, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54, 499-513.

SUKHATME, P.V. et SUKHATME, B.V. (1970). *Sampling Theory of Surveys with Applications*, (2^e éd.). Ames, Iowa: Iowa State University Press.

YATES, F. (1960). *Sampling Methods for Censuses and Surveys*, (3^e éd.). Londres: Griffin.

YATES, F. et GRUNDY, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society, Sér. B*, 15, 253-261.

Tableau 2
Efficacité relative en pourcentage des stratégies 1 et 3 par rapport à la stratégie 2, avec les populations décrites au tableau 1 (n = 4)

Population	Stratégie 1				Stratégie 3
	m = 3	4	5	6	
Numéro					
1.	130.1	118.7	120.8	124.5	127.8
2.	132.6	130.2	—	—	127.1
3.	149.1	—	—	—	147.9
4.	120.7	120.6	122.7	129.7	117.8
5.	129.1	138.7	158.7	—	125.1
6.	158.0	173.1	—	—	139.5
7.	151.9	144.8	169.2	—	131.9
8.	168.5	—	—	—	145.5
9.	118.3	116.3	—	—	109.5
10.	126.6	—	—	—	112.2
11.	113.8	116.2	135.6	129.9	113.8
12.	117.4	128.0	119.0	—	119.3
13.	122.2	120.6	—	—	119.7
14.	124.8	123.1	115.4	113.2	116.3

Il semble donc que, pour que la stratégie 1 soit supérieure à la stratégie 2, il faille choisir *m* de manière que

$$n/2 < m \leq (n - 2). \tag{4.5}$$

Il est toutefois clair que (4.4) n'est qu'une condition suffisante mais non nécessaire. Dans le cas où $n > 6$, la condition (4.5) donne plusieurs choix pour la valeur de *m*, tandis que si $n = 6$, (4.5) implique que $m = 3$. Pour $n = 4$, aucune valeur de *m* ne peut satisfaire (4.5). On s'est donc employé à étudier l'efficacité de la stratégie 1, lorsque $n = 4$, avec un certain nombre de populations naturelles pour diverses valeurs de *m* qui ne satisfont pas (4.5). Une description des populations est donnée au tableau 1, tandis que le tableau 2 donne l'efficacité relative de la stratégie 1 par rapport à celle de la stratégie 2 avec les populations décrites dans le tableau 1. Le tableau 2 compare également la performance de l'estimateur H-T utilisé de concert avec le plan d'échantillonnage de Sampford (1967), ce qu'on désigne ici par la stratégie 3, et la performance de la stratégie 2.

Il ressort du tableau 2 que la performance de la stratégie proposée (stratégie 1) se compare avantageusement à celle de la stratégie 3 dans le cas de la plupart des populations étudiées. Bien entendu, les deux stratégies sont toutes deux supérieures à la stratégie 2.

Pour obtenir l'efficacité relative de la stratégie 1, les unités ont été groupées en ne veillant uniquement qu'à ce que la condition (2.1) soit satisfaite. On a également tenté d'utiliser la méthode de Rao et Lanke (1984) pour former les groupes. Cette méthode n'a toutefois pas conduit à une grande efficacité dans tous les cas. D'autres recherches sont nécessaires pour décider du «meilleur» choix des groupes. Pour certaines populations, il a été impossible de former des groupes satisfaisant à la condition (2.1) avec des valeurs élevées de *m*; c'est pour cette raison que l'efficacité relative dans ces cas-là n'apparaît pas dans le tableau 2.

En conclusion, il semble indiquer de faire un bref commentaire au sujet des cas où la taille de l'échantillon voulue, *n*, est impaire. On peut tirer un échantillon avec PSP pour *n* impair en sélectionnant ($n + 1$) unités à l'aide de la méthode proposée, puis en rejetant au hasard une unité. Les expressions pour les π_i et π_{ij} sont également assez simples dans ce cas-là. De

Tableau 1
Description des populations

Numéro de population	Source	N	y	x
1.	Des Raj (1965)	20	Nombre de ménages	Nombre de ménages estimé à l'oeil
2.	Rao (1963)	14	Nombre d'acres de maïs en 1960	Nombre d'acres de maïs en 1958
3.	Cochran (1963, p. 204)	10	Poids des pêches	Poids des pêches estimé à l'oeil
4.	Hanurav (1967)	20	Population en 1967	Population en 1957
5.	Hanurav (1967)	19	Population en 1967	Population en 1957
6.	Hanurav (1967)	16	Population en 1967	Population en 1957
7.	Hanurav (1967)	17	Population en 1967	Population en 1957
8.	Cochran (1963, p. 325)	10	Nombre de personnes par îlot de logements	Nombre de pièces par îlot de logements
9.	Cochran (1963, p. 156, villes 1-16)	16	Population en 1930	Population en 1920
10.	Cochran (1963, p. 156, villes 33-49)	17	Population en 1930	Population en 1920
11.	Samford (1962, p. 61)	35	Nombre d'acres d'avoine en 1957	Nombre d'acres d'avoine en 1947
12.	Sukhatne et Sukhatne (1970, p. 256, cercles 1-20)	20	Nombre d'acres de blé	Nombre de villages
13.	Sukhatne et Sukhatne (1970, p. 256, cercles 21-40)	20	Nombre d'acres de blé	Nombre de villages
14.	Yates (1960, p. 163)	20	Volume de bois d'oeuvre	Volume de bois d'oeuvre estimé à l'oeil

4. COMPARAISON DU PLAN PROPOSÉ ET DE LA STRATÉGIE
AVEC PPT AVEC REMISE

Dans la présente section, nous comparons l'efficacité des deux stratégies suivantes:

- Stratégie 1. Plan d'échantillonnage proposé utilisé avec l'estimateur de Horvitz-Thompson.
- Stratégie 2. Plan d'échantillonnage avec PPT et avec remise utilisé avec l'estimateur habituel.

La stratégie 1 est plus efficace (variance plus petite) que la stratégie 2 si et seulement si

$$\sum_{m=1}^M \sum_{N_i=1}^n \sum_{i=1}^M \pi_{iN_i} (Y_{iN_i} / P_{iN_i} - Y) (Y_{iN_i} / P_{iN_i} - Y) + \sum_{m=1}^M \sum_{N_i=1}^n \sum_{i \neq j}^M \pi_{iN_i} (Y_{iN_i} / P_{iN_i} - Y) (Y_{jN_i} / P_{jN_i} - Y) < 0. \tag{4.1}$$

Après un certain nombre de manipulations algébriques longues mais élémentaires, l'inégalité (4.1) se ramène à ceci

$$\begin{aligned} & - \sum_{N_i=1}^n \sum_{i=1}^M (n / D_i) \sum_{N_i=1}^n (Y_{iN_i} - Y_i P_{iN_i} / P_i)^2 / (P_i - 2 P_{iN_i}) \\ & - n(n-2) \left[\sum_{m=1}^M \left(Y_i / P_i - Y \right)^2 / [(m-2)(m-1)] \right] \\ & - n(m-2)^{-1} \sum_{m=1}^M \{ (2n-m-2) P_i - (n-2)(m-1)^{-1} \} (Y_i / P_i - Y)^2 < 0, \end{aligned}$$

où

$$Y_i = \sum_{N_i=1}^n Y_{iN_i}$$

De toute évidence, (4.2) se vérifie si

$$(i) \quad (2n-m-2) > 0, \text{ et si}$$

$$(ii) \quad P_i > (n-2) / [(m-1)(2n-m-2)]. \tag{4.3}$$

De plus, comme pour la première étape du plan d'échantillonnage nous utilisons la méthode de Midzuno avec des probabilités révisées $\{P_i'\}$, chacun des P_i doit satisfaire la condition (2.1), c'est-à-dire que chaque P_i doit satisfaire

$$P_i > (n-2) / [n(m-1)].$$

Par conséquent, (4.2) se vérifie si

$$m \leq (n-2).$$

$$(4.4)$$

de Durbin (1967), qui est équivalente à celle de Rao (1963) et de Brewer (1963), donne en général de bons résultats, c'est elle qu'on a décidé d'appliquer à l'étape 2.

3. ESTIMATEUR DE LA VARIANCE

Deux estimateurs sans biais bien connus de la variance de \bar{Y} , $Var(\bar{Y})$, ont été élaborés par Horvitz et Thompson (1952) et par Yates et Grundy (1953). Ces deux estimateurs présentent l'inconvénient de prendre parfois des valeurs négatives. Dans la présente section, on examine un estimateur positif de la variance qui met à profit le fait que le plan d'échantillonnage proposé est à deux degrés.

À l'aide d'un résultat Des Raj (1966), un estimateur non biaisé de $Var(\bar{Y})$ prend la forme suivante:

$$V(\bar{Y}) = \sum_{n/2}^{i=1} \pi_i^2 \sum_{n < v}^n \left[\frac{\pi_{i|v|}}{\pi_{i|}} - 1 \right] \left[\frac{\bar{Y}_{i|v|}}{\bar{Y}_{i|}} - \frac{\pi_{i|v|}}{\pi_{i|}} \right]^2$$

$$+ \sum_{n/2}^i \sum_{n/2}^j \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) \left[\frac{\bar{Y}_i}{\bar{Y}_j} - \frac{\pi_i}{\pi_j} \right]^2, \quad (3.1)$$

où

$$\pi_i = n P_i / 2,$$

$$\pi_{ij} = \frac{n(n-2)}{4(m-2)} \{ (P_i + P_j) - 1 / (m-1) \},$$

$$\pi_{i|v|} = 2 p_{i|v|} / P_i,$$

$$\pi_{i|v|} = \frac{2 p_{i|v|} d_{i|v|} (P_i - p_{i|v|} - d_{i|v|})}{d_i P_i (P_i - 2 p_{i|v|})},$$

et

$$\bar{Y}_i = \sum_{n=1}^n y_{i|v|} / \pi_{i|v|}, \quad (3.2)$$

$y_{i|v|}$ étant la valeur de y chez la n -ième unité dans le i -ième groupe. Les deux termes du membre de droite de l'équation (3.1) correspondent à l'estimateur de la variance de Yates-Grundy calculés respectivement pour la méthode de Durbin (étape 2) et pour la méthode de Midzuno (étape 1). Comme l'estimateur de la variance de Yates-Grundy est toujours positif pour chacune de ces deux méthodes d'échantillonnage, il s'ensuit que l'estimateur de la variance de l'équation (3.1) est également positif. Cependant, l'estimateur en (3.1) n'est ni l'estimateur de la variance de Horvitz-Thompson ni l'estimateur de la variance de Yates-Grundy.

Une fois les m groupes formés, la méthode d'échantillonnage proposée comprend les étapes suivantes:

Etape 1. Sélection de $n/2$ groupes parmi les m groupes, au moyen de la méthode d'échantillonnage de Midzuno (1951) avec probabilités $\{P'_i\}$, c'est-à-dire sélection d'un groupe avec probabilité

$$P'_i = [n(m-1)P_i - (n-2)/(2m-n)], \text{ où } P_i = X_i/X,$$

et des $(n/2) - 1$ autres groupes avec probabilités égales et sans remise.

Etape 2. Sélection de deux unités, dans chacun des groupes choisis, suivant l'une quelconque des méthodes avec PSPT, par exemple à l'aide de la méthode de Durbin (1967), c'est-à-dire sélection dans le i -ième groupe choisi ($i = 1, 2, \dots, n/2$) d'une unité avec probabilité

$$P_{i_n|i} = x_{i_n}/X_i$$

et de la deuxième unité avec probabilité révisée

$$P_{i_n|i_v} = x_{i_v} [1/(X_i - 2x_{i_v}) + 1/(X_i - 2x_{i_n})] / D_i$$

$$D_i = [1 + \sum_{j=1}^n x_{i_n}/(X_i - 2x_{i_n})].$$

Avec cette méthode d'échantillonnage, la probabilité de sélection de la i_n -ième unité est évidemment

où

$$\pi_{i_n} = n P_{i_n}$$
$$P_{i_n} = x_{i_n}/X.$$

De plus, les probabilités de sélection d'une paire d'unités sont

$$\pi_{i_n i_v} = \frac{n P_{i_n} P_{i_v} (P_i - P_{i_n} - P_{i_v})}{D_i (P_i - 2 P_{i_n}) (P_i - 2 P_{i_v})} \quad (2.3)$$

et

$$\pi_{i_n i_v} = \frac{n (n-2) P_{i_n} P_{i_v}}{[(m-1)(m-2) P_i P_j] (P_i + P_j) - 1}, \quad i \neq j; i, j = 1, 2, \dots, m. \quad (2.4)$$

Ainsi, nous voyons que le plan proposé est bel et bien un plan d'échantillonnage avec PSPT. Tel qu'il a été mentionné précédemment, on peut utiliser à l'étape 2 de la méthode proposée n'importe quel plan avec PSPT pour sélectionner deux unités. Comme la méthode

ques souhaitables l'avantage de simplifier le processus de sélection et d'estimation, de permettre de calculer un estimateur de la variance non négatif et d'être plus efficaces que les techniques d'échantillonnage avec probabilités proportionnelles à la taille (PPT) et avec remise. Malheureusement, pour des échantillons de taille plus grande que deux, il n'y a pas encore beaucoup de techniques qui satisfont pleinement à toutes ces exigences.

Dans le présent article, on propose un plan d'échantillonnage pour des échantillons de taille arbitraire n où $n > 2$. La technique est plutôt simple tant pour la sélection de l'échantillon que pour l'estimation, étant donné qu'on peut disposer d'expressions compactes pour les π_i . Elle donne également la possibilité de calculer un estimateur positif de la variance de l'estimateur H-T de Y . La performance réalisée par l'estimateur H-T au moyen du plan proposé est comparée à celle de l'estimateur résultant de l'utilisation d'une technique d'échantillonnage avec PPT et avec remise. À partir de cette comparaison, on trouve une condition suffisante simple qui, si elle est satisfaite, assure une performance du nouveau plan supérieure à celle de l'autre technique. Les résultats d'une étude empirique faite avec quelques populations naturelles indiquent que le plan proposé se compare avantageusement à celui élaboré par Sampford (1967).

2. MÉTHODE D'ÉCHANTILLONNAGE

Solent une population de N unités, y la variable d'intérêt et x une variable auxiliaire prise comme mesure de la taille. On suppose que les valeurs de x sont connues pour toutes les unités de la population. On veut maintenant tirer un échantillon de taille n ($n > 2$). Pour commencer, on suppose que n est pair.

On divise la population en m groupes m ($> n/2$) de telle sorte que le i -ième groupe contiennent N_i unités ($N_i > 2$) et $i = 1, 2, \dots, m$) et que, pour chaque groupe,

$$X_i/X > (n - 2)/(n(m - 1)), \quad (2.1)$$

où

$$X_i = \sum_{N_i}^{n=1} x_{in},$$

x_{in} est la valeur de x chez la n -ième unité du i -ième groupe et $X = X_1 + X_2 + \dots + X_m$.

La condition (2.1) est satisfaite si les X_i ($i = 1, 2, \dots, m$) sont presque égaux. Il a été montré que, dans des populations réelles, telles que celles considérées par Rao et Bayless (1969) et d'autres auteurs, cette condition est satisfaite pour plusieurs valeurs de m si les groupes sont formés de manière que leurs tailles X_i respectives sont presque égales entre elles. Rao et Lanke (1984) ont proposé une méthode de groupement des unités dans laquelle N unités sont groupées en R groupes de telle sorte que la valeur totale de chaque groupe, X_i , est presque égale d'un groupe à un autre et que la taille des groupes est soit $[N/R]$ ou $[N/R] + 1$, où $[x]$ est le plus grand nombre entier contenu dans x . On peut également appliquer la méthode de Rao-Lanke pour former les groupes.

Méthode d'échantillonnage avec probabilités de sélection proportionnelles à la taille

A. DEY et A.K. SRIVASTAVA¹

RÉSUMÉ

On propose ici un nouveau plan d'échantillonnage sans remise avec probabilités inégales de sélection de n unités ($n > 2$) dans une population finie. Ce plan assure que les probabilités de sélection sont proportionnelles à la taille. Il offre l'avantage de simplifier le processus de sélection et d'estimation de produire un estimateur de la variance non négatif. On montre que la variance de l'estimateur de Horvitz-Thompson obtenu à l'aide de ce nouveau plan est plus petite que celle des estimateurs produits habituellement selon un plan d'échantillonnage avec probabilités proportionnelles à la taille et avec remise. Le plan proposé donne également de bons résultats par rapport au plan d'échantillonnage sans remise élaboré par Sampford (1967).

MOTS CLÉS: Échantillonnage avec probabilités inégales; estimateur de Horvitz-Thompson.

1. INTRODUCTION

Dans un plan d'échantillonnage sans remise avec probabilités inégales de sélection de n unités à partir d'une population finie comprenant N unités, si π_i désigne la probabilité d'inclusion de la i -ième unité dans l'échantillon, $i = 1, 2, \dots, N$, l'estimateur de Horvitz-Thompson (1952) (l'estimateur H-T) de Y , qui est la valeur totale de la variable d'intérêt, y , dans la population étudiée, s'exprime

$$(1.1) \quad \bar{y} = \sum_{i \in s} (y_i / \pi_i),$$

où y_i est la valeur de y chez la i -ième unité et où la sommation porte sur les unités incluses dans l'échantillon. La variance de \bar{y} est

$$(1.2) \quad \text{Var}(\bar{y}) = \sum_{i=1}^N \sum_{j>i}^N (\pi_i \pi_j - \pi_{ij}) (y_i / \pi_i - y_j / \pi_j)^2$$

où π_{ij} désigne la probabilité d'inclusion du couple d'unités i et j dans l'échantillon ($i \neq j, i, j = 1, 2, \dots, N$).

On peut s'attendre que la variance de \bar{y} sera beaucoup moins grande si on utilise un plan d'échantillonnage qui assure que les π_i sont proportionnels à une mesure donnée de la taille, disons, x_i , pour $i = 1, 2, \dots, N$, où on suppose que les x_i sont presque proportionnels aux y_i . Les plans d'échantillonnage dans lesquels les π_i sont proportionnels à la taille (PSP). Pour une présentation détaillée des techniques d'échantillonnage avec probabilités inégales, y compris les plans d'échantillonnage avec PSP, voir la monographie de Brewer et Hanif (1983).

Les plans d'échantillonnage avec probabilités inégales et sans remise, en général, et les plans d'échantillonnage avec PSP, en particulier, devraient posséder entre autres caractéristi-

¹ A. Dey et A.K. Srivastava, Indian Agricultural Statistics Research Institute, Library Avenue, Nouvelle-Delhi 110012, Inde.

BIBLIOGRAPHIE

- COX, D.R., et HINKLEY, D.W. (1974). *Theoretical Statistics*. London: Chapman and Hall.
- FAY, R.E. (1985). A jackknifed chi-squared test for complex samples. *Journal of the American Statistical Association*, 80, 148-157.
- IMREY, P.B., KOCH, G.G., et STOKES, M.E. (1982). Categorical data analysis: Some reflections on the log-linear model and logistic regression. Part II: Data analysis. *International Statistical Review*, 50, 35-63.
- KOCH, G.G., FREEMAN, D.H., Jr., et FREEMAN, J.L. (1975). Strategies in the multivariate analysis of data from complex surveys. *International Statistical Review*, 43, 59-78.
- KUMAR, S., et RAO, J.N.K. (1984). Régression logistique et analyse de données de l'enquête sur la population active. *Techniques d'enquête*, 10, 62-81.
- KUMAR, S., et RAO, J.N.K. (1986). On smoothed estimates of unemployment rates from labour force survey data. Dans *Small Area Statistics: An International Symposium '85* (éds. R. Platek et M.P. Singh), Ottawa: Carleton University.
- NEYMAN, J. (1949). Contribution to the Theory of the χ^2 test. Dans *Proceedings of the First Berkeley Symposium on Mathematical Statistics and Probability* (éd. J. Neyman), Berkeley: University of California Press, 230-273.
- RAO, J.N.K., et SCOTT, A.J. (1981). The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence in two way tables. *Journal of the American Statistical Association*, 76, 221-230.
- RAO, J.N.K., et SCOTT, A.J. (1984). On chi-squared tests for multivariate contingency tables with cell proportions estimated from survey data. *Annals of Statistics*, 12, 46-60.
- ROBERTS, G.R. (1985). *Contributions to chi-squared tests with survey data*. Thèse de doctorat, Carleton University, Ottawa.
- ROBERTS, G., RAO, J.N.K., et KUMAR, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, 74, 1-12.
- SINGH, A.C. (1985). On optimal asymptotic tests for analysis of categorical data from sample surveys. Document de travail, Division des méthodes d'enquêtes sociales, Statistique Canada.
- SINGH, A.C., et KUMAR, S. (1986). Categorical data analysis for complex surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, (à paraître).

Les valeurs de $\hat{Q}^{(T)}(\theta)$ pour $T = 54$ et 51 et celles de $\hat{Q}^{(T)}(\theta)$ pour les mêmes valeurs de T permettent de tirer des conclusions similaires.

Le tableau I donne les efficacités relatives aux estimations du taux de chômage $1 - v$ pour les 60 cellules correspondant aux trois estimations lissées. Les trois sortes d'estimations lissées sont: les pseudo-EMV, les estimations de $\hat{Q}^{(51)}$ min et celles de $\hat{Q}^{(54)}$ min. Les variances des pseudo-EMV sont tirées de Kumar et Rao (1986) tandis que celles des estimations de $\hat{Q}^{(T)}$ min ont été établis à l'aide des éléments diagonaux de $B \wedge^T B'$ dans (3.4). Comme le soulignent Kumar et Rao (1986) à propos des pseudo-EMV, les estimations lissées obtenues par la méthode de $\hat{Q}^{(T)}$ min produisent des gains d'efficacité considérables par rapport aux estimations d'enquête. Le rapport entre l'efficacité de trace des estimations lissées et l'efficacité de trace des estimations d'enquête est de 17.9, 18.95 et 19.88 respectivement pour la méthode des pseudo-EMV, celle de $\hat{Q}^{(51)}$ min et celle de $\hat{Q}^{(54)}$ min. Ainsi, la méthode de $\hat{Q}^{(T)}$ min produit des estimations lissées légèrement plus efficaces que la méthode des pseudo-EMV. En ce qui a trait à l'efficacité au niveau de la cellule, le tableau I indique que les pseudo-EMV se comparent très avantageusement aux estimations efficaces de $\hat{Q}^{(T)}$ min.

5. CONCLUSION

En ce qui a trait au calcul des pseudo-EMV, la forme usuelle de la fonction de vraisemblance se rapporte à des échantillons aléatoires simples (c.-à-d., échantillonnage multinomial ou échantillonnage multinomial mixte). Les pseudo-EMV produisent effectivement des estimations de paramètres de modèle convergentes sans que l'on ait à estimer la matrice des covariances T . Toutefois, ce genre d'estimations ne sont pas asymptotiquement efficaces dans le cas de données d'échantillons complexes. En revanche, les estimations de $\hat{Q}^{(T)}$ min sont asymptotiquement efficaces en ce qui concerne la catégorie d'estimateurs fondés sur W (les T premières composantes principales du vecteur b des estimations d'enquête). Pour examiner la performance relative obtenue à l'aide du pseudo-EMV et du $\hat{Q}^{(T)}$ min, il serait souhaitable d'effectuer une étude de simulation pour comparer les efficacités. On adapte les estimations aux plans de sondage complexes en utilisant une matrice T . Si T n'est pas mal conditionnée, c'est-à-dire qu'elle n'a aucune valeur propre relativement minime, les estimateurs par les MCP sont parfaitement stables et, par conséquent, asymptotiquement efficaces. Dans ce genre de situation, on constatera souvent qu'il n'y a pas de réduction de dimension pour de faibles valeurs de ϵ et que T coïncide avec I . De plus, les estimations de $\hat{Q}^{(T)}$ min ne perdront pas de leur efficacité par rapport aux estimations par les MCP. Toutefois, en raison des problèmes d'instabilité qui sont si souvent liés aux données qualitatives recoupées, les estimations de $\hat{Q}^{(T)}$ min devraient bien valoir en robustesse les estimations par les MCP.

REMERCIEMENTS

Les recherches du premier auteur ont été appuyées par Statistique Canada et le Conseil de recherches en sciences naturelles et en génie du Canada.

Nous passons maintenant au calcul d'estimations asymptotiquement efficaces. Les estimations de \hat{Q} min et les estimations par les MCP n'ont pas été calculées parce que \hat{F} est singulière. Les estimations $\hat{\theta}$ de $\hat{Q}^{(T)}$ min ont été calculées pour $\epsilon = .005$ et $\epsilon = .01$ en appliquant la méthode itérative de Newton-Raphson et en utilisant $\hat{\theta}$ comme l'estimateur initial de $\hat{\theta}$ dans la résolution de l'équation (3.2). Les valeurs de $\hat{\theta}_T$ et de $\hat{Q}^{(T)}$ ($\hat{\theta}$) (dans ce dernier cas, le terme négatif de (2.6) disparaît) pour $\epsilon = .005$ et $T = 54$ sont les suivantes

$$\hat{\theta}_{54} = (-2.7112, 0.1944, -0.00196, 0.1432)', \text{ et}$$

$$\hat{Q}^{(54)}(\hat{\theta}_{54}) = 63.4737. \tag{4.7}$$

Pour $\epsilon = .01$, $T = 51$, nous avons

$$\hat{\theta}_{51} = (-2.6739, 0.19702, -0.00202, 0.1364)', \text{ et}$$

$$\hat{Q}^{(51)}(\hat{\theta}_{51}) = 55.2518. \tag{4.8}$$

Tableau 1
Efficacité des estimations lissées des taux de chômage par rapport aux estimations d'enquête^a

N° de cellule	$\hat{Q}^{(51)}$	Min	Pseudo-EMV	N° de cellule	$\hat{Q}^{(51)}$	Min	Pseudo-EMV
1	5.87	5.74	5.44	31	9.01	9.32	8.65
2	3.62	3.62	3.28	32	8.76	9.46	10.68
3	3.45	3.55	3.12	33	36.93	42.93	51.59
4	52.45	51.65	43.46	34	51.55	60.23	81.12
5	104.77	114.30	96.21	35	69.76	79.93	98.37
7	5.33	5.14	4.38	36	9.17	11.01	15.07
8	9.36	9.53	8.09	37	3.48	3.01	3.45
9	6.85	7.16	6.70	38	13.74	15.91	18.00
10	25.65	28.40	26.31	39	66.87	80.98	97.30
11	13.34	14.13	17.73	40	154.81	187.73	221.50
12	27.74	30.85	30.85	41	49.14	67.56	80.61
13	8.64	8.84	7.15	42	17.32	21.73	24.98
14	13.84	13.84	12.37	43	8.57	9.28	8.49
15	8.20	8.49	9.47	44	27.42	31.65	30.74
16	23.14	24.09	27.75	45	58.55	70.67	75.72
17	18.20	21.49	21.51	46	94.11	114.13	121.49
18	9.87	11.14	12.51	47	82.12	112.65	108.52
19	15.87	16.03	13.66	48	26.54	39.41	41.22
20	11.44	11.98	12.56	49	4.95	5.37	4.41
21	12.39	12.39	15.53	50	12.11	14.10	11.17
22	24.83	24.83	32.02	51	6.75	8.61	7.50
23	16.43	18.16	21.55	52	8.83	11.45	9.90
24	6.98	7.83	10.06	53	52.64	71.49	61.14
25	7.49	7.74	6.99	55	3.59	3.93	3.03
26	10.33	11.33	12.32	56	7.33	8.96	8.23
27	6.47	7.18	8.69	57	23.50	29.83	22.11
28	125.81	140.57	172.91	58	221.23	294.59	208.77
29	33.88	38.13	52.00	59	6.45	8.82	6.62
30	14.89	15.24	20.43	60	38.90	52.84	41.96

^a Les cellules 6 et 54 ne sont pas incluses parce que les taux de chômage observés correspondants étaient zéro.

$$\log \frac{1 - v_{jt}}{v_{jt}} = \beta_0 + \beta_1 A_j + \beta_2 A_j^2 + \beta_3 E_j \quad (4.1)$$

où A_j représente le point médian $12 + 5j$ pour le j -ième groupe d'âge ($j = 1, \dots, 10$), et E_j ($j = 1, \dots, 6$) représente le nombre médian d'années de scolarité pour chaque catégorie, soit 7, 10, 12, 13, 14 et 16.

On peut exprimer le modèle (4.1) au moyen de la notation définie à la section 2 en numérotant les soixante cellules par ordre lexicographique. Ainsi, (4.1) peut être écrite $h(v) = X\theta$, où v est le vecteur des taux d'emploi, h est la fonction logit, X est une matrice 60×4 dont la i -ième ligne est $(1, A_i, A_i^2, E_i)$, et θ est $(\beta_0, \beta_1, \beta_2, \beta_3)'$. Nous avons aussi

$$H = (\partial h / \partial v) = D_v^{-1} D_1^{-1} v, B = H^{-1} X, \quad (4.2)$$

où D_v et $D_1^{-1} v$ sont des matrices diagonales dont les éléments de la diagonale sont définis par les indices inférieurs.

À l'aide de la pseudo-fonction de vraisemblance appliquée à la distribution binomiale mixte, Kumar et Rao (1984) ont calculé les pseudo-EMV de θ pour le modèle (4.1) et ont obtenu les valeurs suivantes:

$$\bar{\theta} = (-3.10, 0.211, -0.00218, 0.1509)', \quad (4.3)$$

Ils ont également calculé la valeur du X^2 corrigé de premier ordre de Rao-Scott (qu'ils désignent par G^2); la valeur obtenue, 55.3, entraîne l'acceptation du modèle (2.1) quand on le compare à la distribution X_{56}^2 .

Le modèle (4.1) a aussi été testé à l'aide de la méthode $\bar{Q}^{(T)}$ (voir Singh 1985, Singh et Kumar 1986); là encore, le test s'est soldé par l'acceptation du modèle. Pour $\epsilon = .01$, nous avons $T = 51$ selon la matrice des covariances estimées \bar{T} , qui a été calculée par Kumar et Rao (1984). Si nous prenons maintenant les pseudo-EMV θ , nous avons

$$\bar{Q}^{(51)}(\bar{\theta}) = 58.665 - 4.454 = 54.211 \quad (4.4)$$

Lorsque $\epsilon = .005$, T est égal à 54 et

$$\bar{Q}^{(54)}(\bar{\theta}) = 67.774 - 2.343 = 65.431 \quad (4.5)$$

Lorsque $\epsilon = 0$, $T = 58$ parce que deux cellules renferment des taux de chômage nuls. Dans ce cas,

$$\bar{Q}^{(58)}(\bar{\theta}) = 87.302 - 0.812 = 86.49 \quad (4.6)$$

En comparant les valeurs de $\bar{Q}^{(51)}$, $\bar{Q}^{(54)}$ et $\bar{Q}^{(58)}$ aux distributions X_{47}^2 , X_{50}^2 et X_{54}^2 respectivement, nous constatons que les deux premiers tests entraînent l'acceptation de (4.1) tandis que le dernier entraîne son rejet. Il est possible de vérifier le degré d'instabilité des estimateurs en examinant l'écart qui existe entre $\bar{Q}^{(58)}$ et $\bar{Q}^{(T)}$ (pour $T = 51, 54$); cet écart peut être jugé très significatif en se référant à la distribution X_{58-T}^2 . L'analyse révèle une certaine instabilité pour le test \bar{Q} où il n'y a aucune réduction de dimension. Il est clair que le test des MCP renfermerait aussi des problèmes d'instabilité à cause de la difficulté que l'on aurait à inverser la matrice \bar{T} qui est singulière. La méthode de $\bar{Q}^{(T)}$ min serait donc préférable à la méthode de \bar{Q} min ou à celle des MCP. Dans le but de minimiser la perte d'information, il est recommandé d'utiliser la méthode qui a la valeur de T la plus élevée à la condition, bien sûr, que la valeur de $\bar{Q}^{(T)}$ correspondante entraîne l'acceptation du modèle.

où $B = \partial v / \partial \theta$ et v ont tous deux rapport avec θ . Une méthode itérative comme celle de Newton-Raphson peut servir à résoudre l'équation (3.2). La valeur θ peut être estimée par la méthode des moindres carrés pondérés (MCP) ou la méthode des pseudo-EMV. Il est alors possible de calculer l'estimateur $\hat{Q}_{(T)}^*$ min de v à l'aide de la formule suivante

$$(3.3) \quad \hat{v} = h^{-1}(X\hat{\theta}).$$

Les comportements asymptotiques de $\hat{\theta}$ et de \hat{v} sont définis par la proposition suivante.

Proposition 3.1 Désignons l'expression $(B' \Delta_T B)^{-1}$ par le symbole Δ_T . Nous avons

$$(3.4) \quad \begin{aligned} (a) \quad \hat{\theta} - \theta &\approx \Delta_T B' \Delta_T (\hat{v} - v(\theta)) \sim MVN(0, \Delta_T B') \\ (b) \quad \hat{v} - v &\approx B \Delta_T B' \Delta_T (\hat{v} - v(\theta)) \sim MVN(0, B \Delta_T B') \end{aligned}$$

où " \approx " signifie qu'il y a très peu de possibilité que la différence entre les deux membres de l'expression soient importante.

Cette proposition se démontre par l'application de la méthode- δ aux fonctions $B' \Delta_T (\hat{v} - v(\theta))$ et $\hat{v} - v(\theta)$, ce qui donne

$$\begin{aligned} B' \Delta_T (\hat{v} - v(\theta)) - (B' \Delta_T B) (\hat{\theta} - \theta) &= o_p(1), \\ \hat{v} - v(\theta) - B(\hat{\theta} - \theta) &= o_p(1). \end{aligned}$$

La proposition ci-dessus nous amène à conclure que la matrice des covariances asymptotiques de l'estimateur $\hat{Q}_{(T)}^*$ min $\hat{\theta}$ est l'inverse de la matrice d'information $B' \Delta_T B$ pour θ , qui a été déterminée à l'aide de la fonction de vraisemblance approximative de θ définie en

(2.4). Par conséquent, lorsqu'il n'y a pas de réduction de dimension, l'estimateur $\hat{\theta}$ sera asymptotiquement équivalent à l'estimateur par les MCP de Koch, Freeman et Freeman (1975). Comme nous l'avons dit dans l'introduction, l'estimateur par les MCP a habituellement un comportement instable dans le cas des échantillons finis à cause d'une estimation inefficace de T . Par contre, pour un $\hat{\theta}$ donné ($\epsilon > 0$), l'estimateur $\hat{\theta}$ a normalement un comportement stable avec les échantillons finis, en ce sens que son comportement asymptotique peut fournir une bonne approximation de son comportement en général. Cette condition ne peut être réalisée sans compromettre l'optimalité asymptotique de $\hat{\theta}$ car, pour que cette condition se réalise, cet estimateur doit être limité à une catégorie moindre, c'est-à-dire à la catégorie des estimateurs fondés sur les T premières composantes principales W . En revanche, l'estimateur par les MCP conserve son optimalité asymptotique dans une catégorie plus large, notamment celle des estimateurs fondés sur le vecteur complet de données \hat{v} . Si, pour une valeur faible de ϵ , la statistique $\hat{Q}_{(T)}^*$ donne un résultat non significatif pour H_0 , l'estimateur $\hat{Q}_{(T)}^*$ min correspondant (\hat{v}) est susceptible de se comparer avantageusement à l'estimateur par les MCP au point de vue de la robustesse.

4. ESTIMATIONS DES TAUX DE CHÔMAGE PAR LA MÉTHODE DE $\hat{Q}_{(T)}^*$ MIN

Considérons les données de l'enquête sur la population active (EPA) d'octobre 1980 que Kumar et Rao (1984, 1986) et Roberts, Rao et Kumar (1987) ont analysées en appliquant l'extension de la méthode du X^2 corrigé de Rao-Scott à la régression logistique. Ces auteurs ont montré que le modèle logit défini ci-dessus s'ajustait bien aux estimations d'enquête de taux d'emploi (v_{jt}) pour le tableau de 60 cellules formées simultanément en fonction de l'âge (10 catégories) et du niveau d'instruction (6 catégories). Le modèle est

la meilleure forme de réduction de dimension qui soit, avec une perte minimum d'information. Le test $\tilde{Q}^{(T)}$ (pour de faibles valeurs de ϵ) devrait donc être robuste par rapport au test \tilde{Q} (au sens d'une probabilité d'erreur de première espèce gonflée) dans le cas des échantillons finis à cause de la quasi-singularité possible de Γ . Le test $\tilde{Q}^{(T)}$ est censé atténuer ce problème d'instabilité mais cela ne peut se faire sans qu'il y ait une perte d'information; \tilde{Q} peut donc renfermer des composantes peu fiables dans la direction des vecteurs propres correspondant aux valeurs propres relativement minimes. Cette perte d'information entraîne une diminution de la puissance du test $\tilde{Q}^{(T)}$ pour des alternatives en direction de la (quasi-) singularité. Cependant, cette perte de puissance est compensée par un gain au niveau du contrôle de l'erreur de première espèce. Comme H_0 est un sous-ensemble de H_0^* , $\tilde{Q}^{(T)}$ sera un test conservatif pour H_0 ; cela sert à assurer le contrôle de l'instabilité.

Une version spéciale de $\tilde{Q}^{(T)}(\theta^*)$, asymptotiquement équivalente et décrite par une formule plus simple et semblable à celle du critère X^2 standard de Pearson-Fisher, peut être obtenue en remplaçant θ^* par un estimateur $\hat{\theta}$ qui minimise l'expression $(\hat{v} - v(\theta))' \Delta_T (\hat{v} - v(\theta))$. Nous avons donc,

$$\begin{aligned} \tilde{Q}^{(T)}(\hat{\theta}) &= Y(\hat{\theta})' \Delta_T Y(\hat{\theta}) \\ &= \sum_{i=1}^I [P_i'(\hat{v} - v(\hat{\theta}))]^2 / \lambda_i \\ &\sim \chi_{I-r}^2 \end{aligned} \quad (2.7)$$

Supposons maintenant que le test $\tilde{Q}^{(T)}$ ou un autre test, comme le test X^2 corrigé, nous a amené à croire qu'un modèle H_0 convient à un vecteur de données \hat{v} . Dans la section suivante, nous définissons une méthode asymptotiquement efficace pour estimer les paramètres H_0 , utilisant la statistique $\tilde{Q}^{(T)}$. Les estimations ainsi obtenues permettent d'établir une série d'estimations lissées de v qui correspondent aux estimations d'enquête \hat{v} .

3. L'ESTIMATEUR $\tilde{Q}^{(T)}$ MIN

Considérons la fonction de vraisemblance approximative pour la moyenne μ des T premières composantes principales W de \hat{v} , qui sont définies par l'équation (2.4). Supposons que le modèle $H_0: h(v) = X\theta$ est accepté. Alors, la fonction du noyau $K(\theta)$ de la fonction de

$$K(\theta) = (W - \mu(\theta))' D^{-1} (W - \mu(\theta))$$

$$= (\hat{v} - v(\theta))' \Delta_T (\hat{v} - v(\theta)) \quad (3.1)$$

La valeur $\hat{\theta}$ qui minimise $K(\theta)$ correspond à l'EMV de θ pour la fonction de vraisemblance approximative de μ sous H_0 . Cet estimateur $\hat{\theta}$ sera asymptotiquement efficace (ou le meilleur estimateur asymptotiquement normal (MEAN) au sens de Neyman, 1949) dans une catégorie restreinte, notamment dans la catégorie de tests fondés sur W . D'après l'estimateur de X^2 min de Neyman (1949), l'estimateur $\hat{\theta}$ est donc devenu dans Singh (1985) l'estimateur de dimension ϵ par Δ_T . Par conséquent, $\hat{\theta}$ varie avec ϵ .

Etant donné les composantes principales W et sous H_0 , nous pouvons calculer les estimations lissées de v de la façon suivante. Nous devons trouver tout d'abord la valeur $\hat{\theta}$ qui minimise $K(\theta)$; en d'autres termes, la valeur recherchée est la solution de r équations

Choisissons maintenant un faible facteur de réduction de dimension $\epsilon (>0)$ (.01 ou .005 peuvent être des valeurs pratiques d' ϵ). Trouvons un nombre T de telle sorte qu'avec les valeurs propres $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_l$ de la matrice des covariances estimées $\hat{\Gamma}$, nous ayons

(2.3)
$$T = \max \left\{ t : t > r \text{ et } \sum_{i=1}^t \lambda_i / \sum_{i=1}^l \lambda_i \geq \epsilon \right\}.$$

Bien qu'aléatoire, la variable T peut être considérée comme indépendante pour nos fonctions asymptotiques. Il convient de souligner qu'en l'absence de valeurs propres relativement minimales (c.-à-d. que $\hat{\Gamma}$ n'est pas mal conditionnée), une faible valeur de ϵ n'aura normale-ment pas d'effet de réduction de dimension et T coïncidera alors avec l .

Examinons le problème qui consiste à tester l'hypothèse nulle H_0 par rapport à l'hypothèse alternative $K_0 : h(v) \neq X\theta$ dans la catégorie de tests fondés sur les T premières composantes principales W de \hat{v} . Soit P_i le vecteur propre normé correspondant à λ_i (ce vecteur n'est pas nécessairement unique), et soit M_T la matrice $(l \times T)$ des vecteurs pro-pres P_i correspondant aux T premières valeurs propres les plus élevées. Alors,

(2.4)
$$W = M_T^T \hat{v} \sim MVN(\mu, D_T/n),$$

où
$$\mu = M_T^T v, D_T = \text{diag}(\lambda_1, \dots, \lambda_T).$$

Lorsqu'il est fondé sur les composantes principales W , le test d'hypothèse qui portait à l'origine sur un vecteur v de dimension l se ramène à un test d'hypothèse sur un paramètre μ de dimension T , où l'hypothèse est définie comme suit:

(2.5)
$$H_0: \mu = M_T^T h^{-1}(X\theta) \text{ vs } K_0: \mu \neq M_T^T h^{-1}(X\theta).$$

On peut calculer la statistique $\tilde{Q}^{(T)}$ comme une cote des composantes principales utilisant la fonction de vraisemblance approximative de θ , qui est définie par la distribution limite de W (2.4) pour le calcul des cotes efficaces (voir Cox et Hinkley 1974, p. 321-324). Nous allons désigner $\tilde{Q}^{(T)}$ comme un test de cote généralisée en vertu duquel H_0 serait rejetée pour les valeurs fortes de la formule quadratique

(2.6)
$$\tilde{Q}^{(T)}(\theta^o) = Y(\theta^o)' \Delta_T Y(\theta^o) - Z_T(\theta^o)' \wedge_T Z_T(\theta^o) \sim \chi^2_{T-r}$$

où
$$Y(\theta^o) = \hat{v} - v(\theta^o), \Delta_T = n \sum_{i=1}^T (P_i^T P_i^T / \lambda_i), \wedge_T = (B^T \Delta_T B)^{-1},$$

 $Z_T(\theta^o) = B^T \Delta_T Y(\theta^o), B = (\partial v / \partial \theta), \wedge_T = (B^T \Delta_T B)^{-1},$
et θ^o est un point fixe dans l'espace des paramètres nuls. Dans le calcul de $\tilde{Q}^{(T)}$, n'importe quelle estimation racine n -convergente de θ , sous H_0 , peut être substituée à θ^o , comme peut l'être la pseudo-EMV de θ . Il est à noter que $\tilde{Q}^{(T)}$ dans (2.6) est en réalité définie par une formule quadratique dans W mais que, pour des raisons de commodité, elle est exprimée en fonction de \hat{v} .

Pour tester H_0 par rapport à K_0 dans la catégorie de tests fondés sur W , l'optimalité asymptotique du test $\tilde{Q}^{(T)}$ provient de celle de la cote. Pour les faibles valeurs de $\epsilon > 0$, \hat{v} et W seront comparables en ce sens que les composantes principales, dans ce cas, assureront

binomiaux indépendants pour estimer les paramètres d'un modèle logit après avoir vérifié la validité de l'ajustement de ce modèle pour les données de l'EPA d'octobre 1980. Ils ont constaté que les estimations lissées de taux de chômage étaient beaucoup plus efficaces que les estimations d'enquête dans l'exemple de l'EPA.

Les pseudo-EMV sont particulièrement utiles lorsqu'il est impossible ou difficile de calculer la fonction de vraisemblance à cause de la complexité du plan de sondage. Dans des conditions de régularité acceptables, la méthode des pseudo-EMV produit des estimations convergentes et asymptotiquement normales (Imrey, Koch et Stokes 1982). Dans cet article, nous tentons de produire des estimations asymptotiquement efficaces (au sens défini à la section 3) des paramètres de modèle et, partant, des estimations de domaines. Nous décrivons aussi l'estimateur $\hat{Q}^{(T)}$ min, qui a été défini dans Singh (1985), à l'aide de la méthode des cotes généralisées; on peut considérer cet estimateur comme analogue à l'estimateur X^2 min de Neyman pour les échantillons aléatoires simples. Il convient de souligner que la méthode des MCP (moindres carrés pondérés) produit également des estimateurs asymptotiquement efficaces dans le cas des plans de sondage complexes (Koch, Freeman et Freeman 1975). Toutefois, ces estimateurs sont habituellement instables avec des échantillons de taille moyenne à cause de la quasi-singularité de la matrice des covariances estimées des estimations de cellules (voir Imrey, Koch et Stokes 1982; Fay 1985). Par contre, les estimateurs $\hat{Q}^{(T)}$ min sont à l'abri de cette instabilité. Nous verrons que la méthode de $\hat{Q}^{(T)}$ min peut permettre de contourner le problème de l'instabilité en utilisant une version modifiée de la matrice des covariances estimées, où l'on élimine les valeurs propres relativement minimes de la décomposition spectrale de la matrice.

La section 2 présente la notation des termes utilisés dans l'étude ainsi qu'une brève analyse du test $\hat{Q}^{(T)}$. Dans la section 3, nous décrivons l'estimateur $\hat{Q}^{(T)}$ min et son comportement asymptotique. La section 4 contient l'exemple d'application aux données de l'EPA. Cet exemple nous permet de constater avec intérêt que les pseudo-EMV sont presque aussi efficaces que les estimations efficaces $\hat{Q}^{(T)}$ min au niveau de la cellule. Si nous prenons une mesure globale, comme l'efficacité de trace, nous constatons que les pseudo-EMV ne sont que légèrement inférieures aux estimations $\hat{Q}^{(T)}$ min. Enfin, la section 5 sert de conclusion.

2. LE TEST $\hat{Q}^{(T)}$; BRÈVE ANALYSE

Nous allons décrire brièvement le test $\hat{Q}^{(T)}$ afin de motiver l'utilisation de la méthode d'estimation de $\hat{Q}^{(T)}$ min (pour plus de détails, voir Singh 1985, Singh et Kumar 1986). Soit I le nombre de domaines disjoints et v_i le paramètre d'intérêt pour le i -ième domaine. Considérons un modèle pour $v = (v_1, v_2, \dots, v_I)'$ tel que

(2.1) $H_0: h(v) = X\theta$

où X est une matrice connue $I \times r$ de plein rang r , θ est un r -vecteur de paramètres inconnus et h est une fonction univariante continûment différentiable, par exemple log ou logit. Posons \hat{v} comme I -vecteur des estimations d'enquête et supposons qu'en vertu du théorème de la limite centrale approprié,

(2.2) $\hat{v} \sim MVN(v, I/n)$

où " \sim " signifie "distribué asymptotiquement comme", n est la taille de l'échantillon et I est la matrice des covariances asymptotiques de $\sqrt{n}(\hat{v} - v)$.

Sur l'estimation efficace des taux de chômage à l'aide de données de l'enquête sur la population active

S. KUMAR et A.C. SINGH¹

RÉSUMÉ

Tout comme avec la méthode d'estimation de X^2 min de Neyman (1949) pour les échantillons aléatoires simples, la méthode de $\bar{Q}^{(T)}$ minimum proposée par Singh (1985) pour les plans de sondage complexes produit des estimateurs de paramètres de modèle asymptotiquement efficaces. La variable $\bar{Q}^{(T)}$ peut être considérée comme une variable X^2 pour des données d'enquête qualitatives. Les estimateurs $\bar{Q}^{(T)}$ min peuvent remplacer avantageusement les estimateurs obtenus par les moindres carrés pondérés qui ont souvent un comportement instable dans le cas des échantillons complexes. Les auteurs exposent tout d'abord la méthode de $\bar{Q}^{(T)}$ min puis l'appliquent à l'estimation des paramètres d'un modèle logit pour des estimations de taux de chômage calculées à partir des données de l'EPA d'octobre 1980 classées selon deux critères: l'âge et le niveau d'instruction. Il apparaît que l'efficacité de trace des estimations lissées calculées par Kumar et Rao (1986) à l'aide de la méthode des pseudo-EMV peut être légèrement améliorée à l'aide de la méthode de $\bar{Q}^{(T)}$ min. Détail digne de mention, les pseudo-EMV pour les cellules prises individuellement ont un comportement très comparable à celui des estimateurs efficaces $\bar{Q}^{(T)}$ min utilisés pour l'EPA.

MOTS CLÉS: Pseudo-EMV; estimateur par les moindres carrés pondérés (MCP); estimateur $\bar{Q}^{(T)}$ min; efficacité asymptotique; fonction de vraisemblance approximative; cote généralisée.

1. INTRODUCTION

En se basant sur les données de l'enquête sur la population active (EPA) d'octobre 1980, Kumar et Rao (1984, 1986) ont proposé et analysé un modèle de régression logistique (modèle logit) pour les taux de chômage. Ils ont utilisé la théorie élaborée par Roberts (1985) et Roberts, Rao et Kumar (1987). Ces derniers ont généralisé la méthode proposée par Rao et Scott (1981, 1984) pour corriger le X^2 en fonction des effets du plan de sondage et ont vérifié la validité de l'ajustement du modèle logit. Kumar et Rao ont considéré des taux de chômage pour diverses cellules (ou domaines) obtenues en classant simultanément la population visée en fonction de deux critères: l'âge et le niveau d'instruction. Le modèle logit renfermait des effets linéaires et quadratiques pour l'âge tandis qu'il ne renfermait que des effets linéaires pour le niveau d'instruction. Singh et Kumar (1986) ont analysé les mêmes données d'enquête au moyen d'une autre méthode, le test $\bar{Q}^{(T)}$, qui a été proposée par Singh (1985). Le test $\bar{Q}^{(T)}$ est un test du chi-carré fondé sur une cote généralisée de composantes principales. Ce test a produit des résultats comparables à ceux obtenus par la méthode du X^2 corrigé.

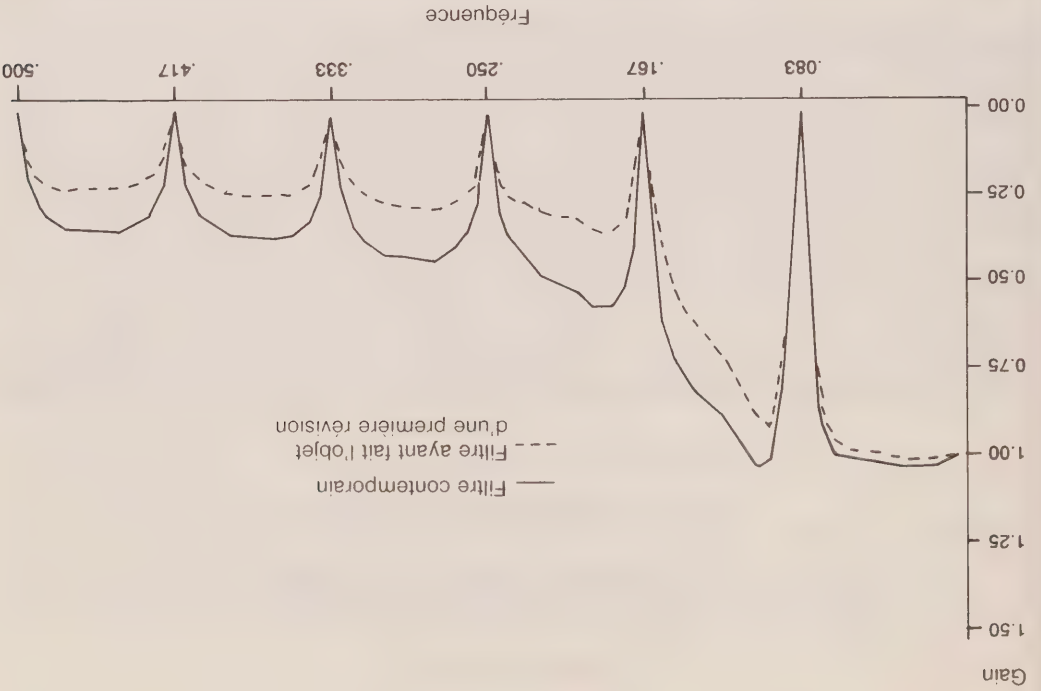
Dès qu'on a construit un modèle convenable, il faut trouver de bons estimateurs de ces paramètres. Ces estimateurs permettent ensuite d'estimer assez fidèlement les taux réels des domaines. Les estimations ainsi obtenues, souvent appelées "estimations lissées", sont particulièrement utiles dans les cas où les estimations d'enquête ne sont pas assez précises à cause d'un nombre insuffisant d'observations. Comme les estimations lissées sont calculées après qu'on a vérifié la validité de l'ajustement du modèle, il faut s'attendre que le biais des estimations soit négligeable. Kumar et Rao (1986) ont utilisé la méthode des pseudo-EMV (pseudo-estimateurs du maximum de vraisemblance) dans une situation fictive d'échantillons

¹ S. Kumar, méthodologiste principal, Division des méthodes d'enquêtes sociales, Statistique Canada, 4-D2, Immeuble Jean Talon, Parc Tunney, Ottawa (Ontario), K1A 0T6. A.C. Singh, Département de mathématique et de statistique, Memorial University of Newfoundland, St. John's (Terre-Neuve), Canada, A1C 5S7.

- BAYER, A., et WILCOX, D. (1981). An evaluation of concurrent seasonal adjustment. Document technique, Board of Governors of the Federal Reserve System.
- DAGUM, E.B. (1978). *Comparison and Assessment of Seasonal Adjustment Methods for Labour Force Series*. Stock No. 052-003-00603-1, U.S. Government Printing Office.
- DAGUM, E.B. (1980). *La méthode de désaisonnalisation X-11-ARMMI*. N° 12-564F au répertoire, Statistique Canada.
- DAGUM, E.B. (1982a). Revisions of time varying seasonal filters. *Journal of Forecasting*, 1, 173-187.
- DAGUM, E.B. (1982b). The effects of asymmetric filters on seasonal factor revisions. *Journal of the American Statistical Association*, 77, 732-738.
- DAGUM, E.B. (1982c). Revisions of seasonally adjusted data due to filter changes. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 39-45.
- DAGUM, E.B., et MORRY, M. (1984). Basic issues on the seasonal adjustment of the Canadian Consumer Price Index. *Journal of Business and Economic Statistics*, 2, 250-259.
- DAGUM, E.B. (1987). Monthly versus annual revisions of concurrent seasonally adjusted series. Dans *Time Series and Econometric Modelling*, (eds. I.B. MacNeill et G.J. Umphrey), New York: D. Reidel, 131-196.
- DAGUM E.B., et LANIET, N. (1987). Revisions of trend-cycle estimators of moving average seasonal adjustment method. *Journal of Business and Economic Statistics*, (en voie de rédaction).
- KENNY, P., et DURBIN, J. (1982). Local trend estimation and seasonal adjustment of economic time series. *Journal of the Royal Statistical Society, Ser. A*, 145, 1-41.
- MARAVALL, A. (1986). An application of model-based estimation of unobserved components. *International Journal of Forecasting*, 2, 305-318.
- MOORE, G.H., BOX, G.E.P., KAITZ, H.B., STEPHENSON, J.A., et ZELLNER, A. (1981). Seasonal adjustment of the monetary aggregates. Dans *Report of the Committee of Experts on Seasonal Adjustment Techniques*, Washington: Board of Governors of the Federal Reserve System.
- McKENZIE, S. (1984). Concurrent seasonal adjustment with Census X-11. *Journal of Business and Economic Statistics*, 2, 235-249.
- PIERCE, D.A. (1980). Data revisions with moving average seasonal adjustment procedures. *Journal of Econometrics*, 14, 95-114.
- PIERCE, D., et McKENZIE, S. (1985). On concurrent seasonal adjustment. Special Studies Paper 164, Federal Reserve Board.

BIBLIOGRAPHIE

Figure 3a. Fonctions de gain du filtre contemporain et du filtre ayant fait l'objet d'une première révision du X-11-ARMMI avec extrapolations ARMMI ($\theta = .40$, $\Theta = .60$).



Déphasage

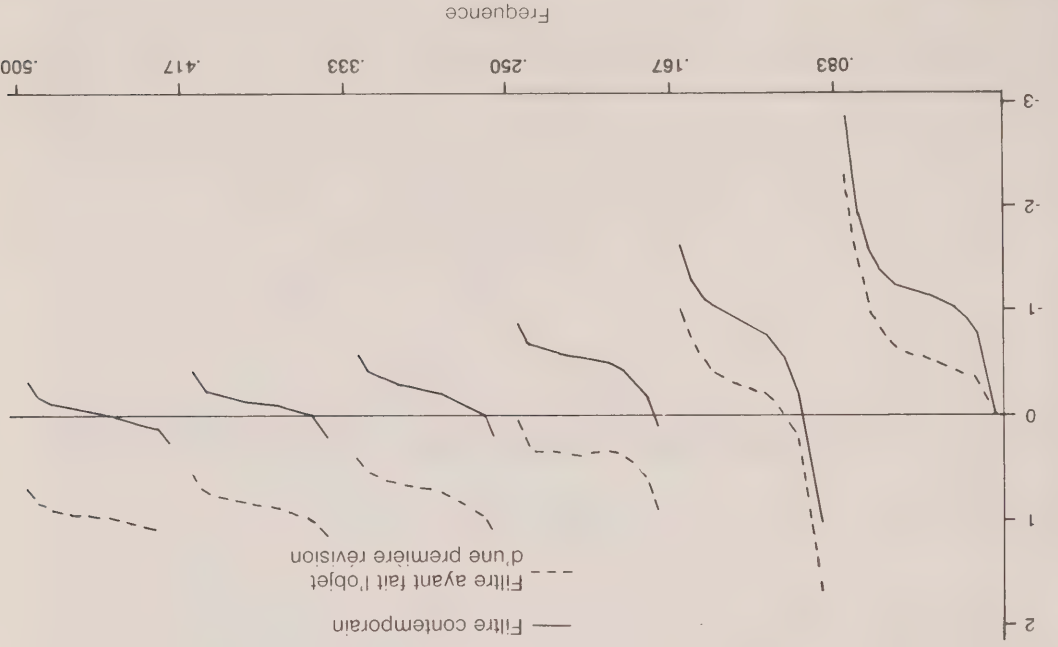


Figure 3b. Fonctions de déphasage du filtre contemporain et du filtre ayant fait l'objet d'une première révision du X-11-ARMMI avec extrapolations ARMMI ($\theta = .40$, $\Theta = .60$).

Figure 2a. Fonctions de gain du filtre contemporain et du filtre ayant fait l'objet d'une première révision du X-11-ARMMI sans extrapolations ARMMI.

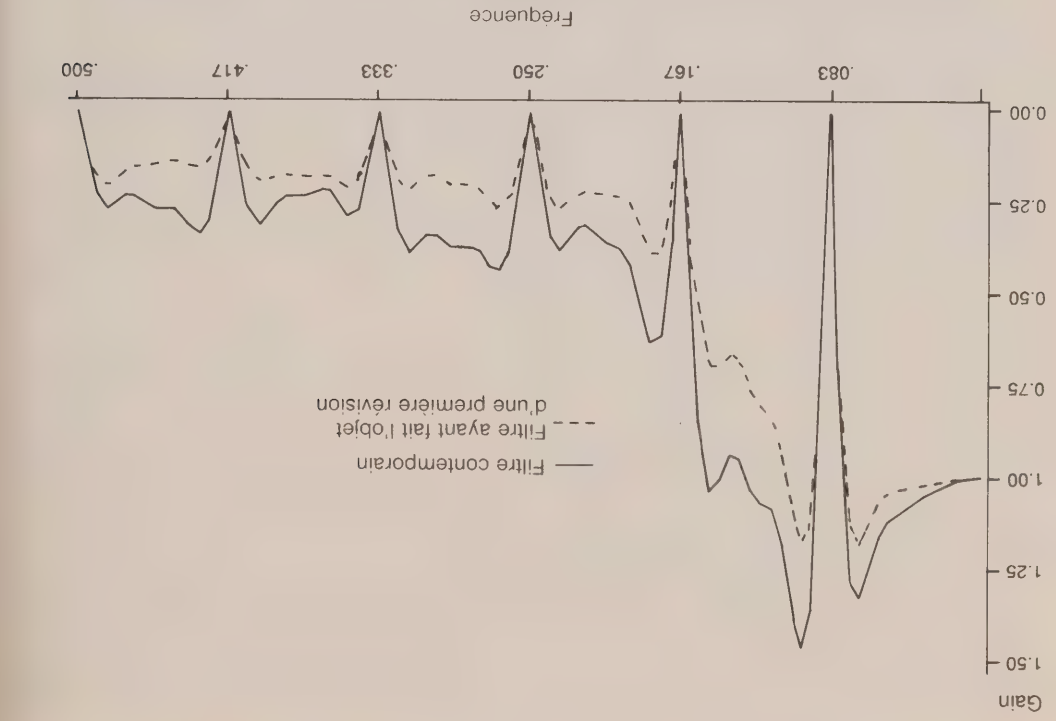


Figure 2b. Fonctions de déphasage du filtre contemporain et du filtre ayant fait l'objet d'une première révision du X-11-ARMMI sans extrapolations ARMMI.

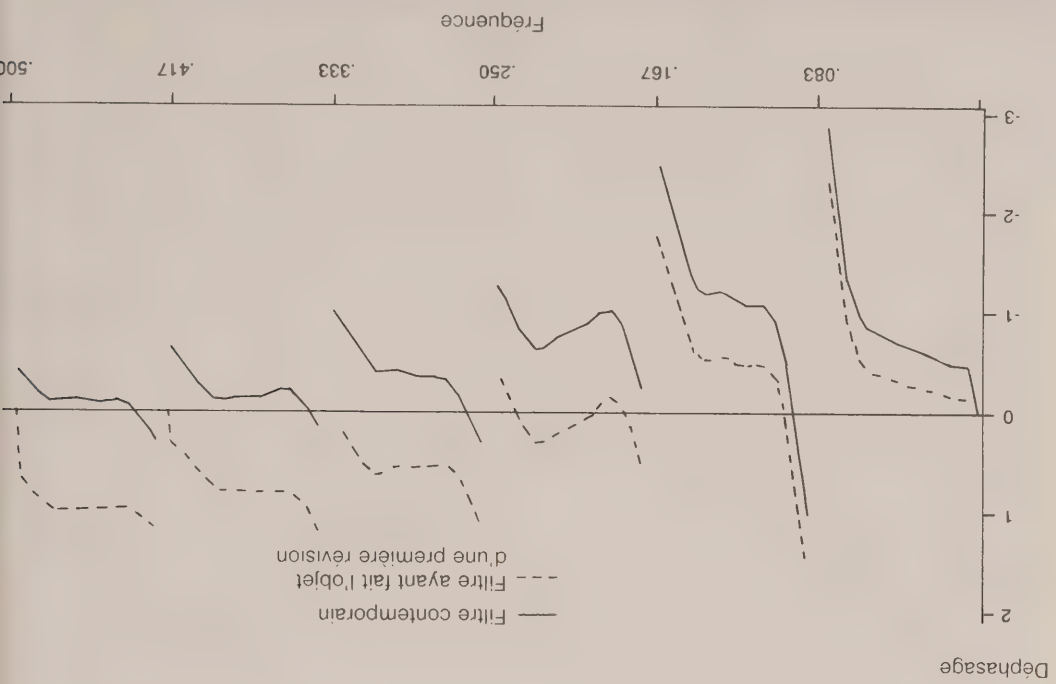


Figure 1. Fonctions de gain du filtre symétrique de la tendance-cycle et du filtre de désaisonnalisation du X-11-ARMMI.



Tableau 3
Evolution chronologique des révisions totales apportées par les filtres de la tendance-cycle et les filtres asymétriques de désaisonnalisation selon la méthode X-11-ARMMI

Révisions $R_{(k,k)*}$	Sans extrapolations		Avec extrapolations tirées d'un modèle $(0,1,1) (0,1,1)^2$ $\theta = .40 \quad \Theta = .60$	
	Filtres de la tendance-cycle	Filtres de désaisonnalisation	Filtres de la tendance-cycle	Filtres de désaisonnalisation
$R_{(48,0)}$.45	.36	.41	.32
$R_{(48,1)}$.27	.33	.26	.32
$R_{(48,2)}$.15	.32	.15	.32
$R_{(48,3)}$.11	.32	.11	.31
$R_{(48,4)}$.12	.32	.11	.31
.
$R_{(48,12)}$.10	.23	.09	.20
$R_{(48,24)}$.07	.13	.05	.10
$R_{(48,36)}$.03	.05	.02	.04
$R_{(48,47)}$.01	.01	.01	.01

* $\ell = 48$ pour le filtre de la tendance-cycle "final" et $\ell = 42$ pour le filtre de désaisonnalisation final. Toutefois, les valeurs des révisions apportées par les filtres de désaisonnalisation sont calculées aussi pour $\ell = 48$ puisqu'après $\ell = 42$, ces valeurs sont finales et, de ce fait, ne varient pas.

Young (1968) a calculé, pour la variante X-1 de la Census Method II, les filtres linéaires combinés qui sont appliqués à la série initiale pour produire une estimation centrale (symétrique) de la tendance-cycle. Ces filtres sont semblables à celui du X-11-ARMMI avec ou sans extrapolations ARMMI. Dagum et Laniel (1987) ont poussé plus loin les travaux de Young (1968) en calculant l'estimation des filtres asymétriques de la tendance-cycle du X-11-ARMMI avec ou sans extrapolations ARMMI.

La figure 1 montre la fonction de gain des filtres de désaisonnalisation centraux (symétriques) et celle des filtres des données désaisonnalisées lissées (tendance-cycle). Il est clair que les filtres de la tendance-cycle suppriment tout bruit présent dans la série, le bruit étant défini comme la puissance présente dans toutes les fréquences $\omega \leq .166$. Cette dernière fréquence correspond à la première harmonique de la fréquence saisonnière fondamentale d'une série mensuelle. Cette structure découle de la convolution des filtres de désaisonnalisation avec le filtre de Henderson de 13 termes (tendance-cycle).

La figure 2a montre la fonction de gain des filtres appliqués aux facteurs saisonniers contemporains et des filtres de la tendance-cycle ayant fait l'objet d'une première révision mensuelle selon le X-11-ARMMI *sans* extrapolations ARMMI. La figure 2b montre les déphasages correspondants exprimés en mois plutôt qu'en radians. Nous constatons que le gain pour tout $\omega \leq .166$ est beaucoup plus grand dans le cas de ces deux filtres asymétriques que dans le cas du filtre central. Nous observons, en outre, de fortes amplifications pour les fréquences qui sont près de la fréquence saisonnière fondamentale. Toutes ces observations signifient que les valeurs désaisonnalisées contemporaines et les valeurs désaisonnalisées lissées ayant fait l'objet d'une première révision renfermeront plus de bruit que les estimations finales. Par ailleurs, on note que les déphasages sont très peu prononcés, soit moins d'un mois pour les fréquences cycliques les plus importantes, $0 < \omega < .055$ (c.-à-d., cycles d'une périodicité égale ou supérieure à 18 mois).

Les figures 3a et 3b montrent le gain et le déphasage des filtres appliqués aux facteurs saisonniers contemporains et des filtres de la tendance-cycle ayant fait l'objet d'une première révision mensuelle selon la méthode X-11-ARMMI avec extrapolations ARMMI. Les extrapolations sont tirées d'un modèle MMI $(0,1,1)(0,1,1)_{12}$ avec $\theta = .40$ et $\Theta = .60$. Les fonctions de gain de la figure 3a s'apparentent plus à la fonction de gain du filtre symétrique (central) que celles observées pour la méthode X-11-ARMMI sans les extrapolations ARMMI. Il n'y a aucune amplification autour de la fréquence saisonnière fondamentale mais on constate, comme dans le premier cas, une diminution de puissance aux fréquences élevées. En revanche, le déphasage est plus prononcé (près d'un mois) aux fréquences basses et moins prononcé à toutes les fréquences élevées.

Dagum et Laniel (1987) ont étudié l'évolution chronologique des révisions apportées par les filtres de la tendance-cycle et l'ont comparée à celle des révisions apportées par les filtres de désaisonnalisation. Les résultats de leur analyse, qui sont résumés dans le tableau 3, montrent que les révisions totales apportées par les filtres asymétriques de la tendance-cycle tendent vers zéro beaucoup plus rapidement que celles apportées par les filtres de désaisonnalisation correspondants. De fait, la révision totale apportée par le filtre de la tendance-cycle trois mois après l'application du filtre contemporain n'est que de 0.1, alors que le filtre de désaisonnalisation ne produit pas de valeur comparable avant que 24 mois n'aient été ajoutés à la série. Sauf en ce qui concerne les premières révisions totales (filtre appliqué aux facteurs saisonniers courants), les révisions apportées par les filtres de la tendance-cycle sont inférieures à celles apportées par les filtres de désaisonnalisation correspondants. En outre, les premières tendent vers zéro beaucoup plus rapidement que les secondes.

Les séries de flux, c'est-à-dire qui découlent de l'accumulation de données quotidiennes dans les mois civils, sont soumises à un effet systématique attribuable aux variations liées aux jours ouvrables. Ce genre de variations sont imputables principalement au fait que l'activité commerciale fluctue selon les jours de la semaine. Les pratiques comptables et les méthodes de présentation des états financiers peuvent aussi expliquer ce genre de variations. Par exemple, les magasins où la comptabilisation des opérations se fait le vendredi ont tendance à déclarer un chiffre de ventes plus élevé dans les mois qui comptent cinq vendredis que dans ceux qui en comptent quatre. Dans le programme X-11-ARMMI, on estime l'effet des variations liées aux jours ouvrables en appliquant la méthode des moindres carrés ordinaires à un modèle de régression simple déterministe; par conséquent, les poids estimés pour chaque jour varient chaque fois qu'une nouvelle observation s'ajoute à la série. Comme les techniques de régression sont très sensibles aux valeurs aberrantes, ces variations de poids sont parfois inutilement élevées.

Lorsque les séries sont soumises à une désaisonnalisation contemporaine, les estimations des variations liées aux jours ouvrables varient constamment. Pour éviter des révisions inutilisées, Statistique Canada utilise habituellement les poids calculés par le programme à la fin de l'année civile précédente ou ceux fournis par les utilisateurs comme facteurs de pondération préalable pour l'année courante. Par la suite, ces poids sont révisés annuellement. L'effet des variations liées aux jours ouvrables doit être supprimé d'une série avant l'application d'un modèle ARMMI, car ce genre de modèle n'est pas conçu pour tenir compte des variations liées aux jours ouvrables. En d'autres mots, si l'on applique le programme X-11-ARMMI avec extrapolations ARMMI à une série qui est soumise à des variations liées aux jours ouvrables, on estime *a priori* ces variations et, si elles sont appréciables, on les extrait de la série initiale avant l'application du modèle ARMMI.

Un autre problème que soulève la désaisonnalisation contemporaine a trait au nombre de fois que des modèles ARMMI devraient être définis. En règle générale, Statistique Canada utilise l'option automatisée ARMMI une fois par année; si le modèle est accepté, il est maintenu fixe toute une année. Seuls les paramètres varient lorsque des observations s'ajoutent. Pour maintenir fixe le modèle, il convient d'appliquer l'option du modèle fourni par l'utilisateur. En maintenant le modèle ARMMI fixe, on évite les révisions inutilisées qui peuvent découler de changements de modèle attribuables uniquement à l'existence de valeurs aberrantes.

4. LISSAGE DE SÉRIES DÉSASONNALISÉES VOLATILES

Un des principaux objectifs de la désaisonnalisation des séries chronologiques économiques est de produire de l'information sur la conjoncture économique actuelle et, plus particulièrement, de déterminer la phase du cycle dans laquelle se trouve l'économie. Comme la désaisonnalisation consiste à éliminer les variations saisonnières, laissant par le fait même les variations de la tendance-cycle et les variations irrégulières, il est souvent difficile de déterminer la tendance à court terme ou les points de retournement cycliques des séries qui sont très irrégulières. Dans ce cas, il peut être préférable de lisser les séries désaisonnalisées au moyen d'estimateurs de la tendance-cycle, qui éliminent le plus possible la composante irrégulière sans toucher à la composante cyclique.

L'utilisation de valeurs de la tendance-cycle a été analysée par plusieurs auteurs, dont récemment Moore et coll. (1981), Kenny et Durbin (1982), Maravall (1986) et Dagum et Laniel (1987). Bien que cette pratique soit encore peu répandue, certains organismes statistiques tels que Statistique Canada et l'Australian Bureau of Statistics lisent certaines de leurs séries désaisonnalisées, notamment celles qui sont fortement touchées par les irrégulières.

Tableau 1
Révisions de première, seconde et troisième année apportées
par le filtre contemporain du X-11-ARMMI

Révisions $R_{(k)}$	Sans extrapolations	ARMMI	$\theta = .40$	$\theta = .80$	$\theta = .80$
Avec extrapolations ARMMI tirées d'un modèle $(0,1,1) (0,1,1)_{12}$					
$R_{(1,0)}$.12	.12	.12	.12	.06
$R_{(2,0)}$.13	.13	.13	.13	.08
$R_{(3,0)}$.13	.13	.13	.13	.08
$R_{(4,0)}$.13	.13	.13	.13	.09
$R_{(5,0)}$.15	.13	.13	.13	.09
$R_{(6,0)}$.17	.13	.13	.13	.09
$R_{(7,0)}$.16	.13	.13	.13	.09
$R_{(8,0)}$.16	.13	.13	.13	.09
$R_{(9,0)}$.16	.13	.13	.13	.09
$R_{(10,0)}$.16	.14	.14	.14	.09
$R_{(11,0)}$.16	.14	.14	.14	.09
$R_{(12,0)}$.29	.28	.27	.26	.26
$R_{(13,1)}$.27	.27	.27	.26	.26
$R_{(14,2)}$.27	.27	.27	.26	.26
$R_{(23,11)}$.27	.26	.26	.26	.26
$R_{(24,12)}$.20	.16	.17	.16	.16
$R_{(24,13)}$.18	.17	.17	.16	.16
$R_{(36,24)}$.16	.17	.17	.16	.16

Tableau 2
Révisions de première année apportées par le filtre de désaisonnalisation
ayant fait l'objet d'une première révision mensuelle

Révisions $R_{(k)}$	Sans extrapolations	ARMMI	$\theta = .40$	$\theta = .80$	$\theta = .80$
Avec extrapolations ARMMI tirées d'un modèle $(0,1,1) (0,1,1)_{12}$					
$R_{(2,1)}$.07	.07	.10	.10	.06
$R_{(3,1)}$.07	.07	.10	.10	.06
$R_{(4,1)}$.07	.07	.10	.10	.07
$R_{(5,1)}$.08	.08	.10	.10	.08
$R_{(6,1)}$.10	.10	.11	.11	.08
$R_{(7,1)}$.11	.11	.11	.11	.08
$R_{(8,1)}$.11	.11	.11	.11	.08
$R_{(9,1)}$.11	.11	.11	.11	.08
$R_{(10,1)}$.12	.12	.11	.11	.08
$R_{(11,1)}$.12	.12	.12	.12	.08

La forte diminution observée dans les trois premières révisions consécutives s'explique par l'amélioration des poids des filtres de Henderson (tendance-cycle). Le changement de tendance observé consécutivement à $\ell = 12$ et à $\ell = 13$ est dû aux améliorations apportées au filtre saisonnier qui perd de son asymétrie d'une année à l'autre jusqu'à ce que trois années complètes s'ajoutent à la série. On observe les corrections les plus prononcées à $\ell = 1$ et à $\ell = 12$. Étant donné la non-monotonie des révisions mensuelles, il n'est pas recommandé de réviser les estimations de facteurs saisonniers contemporains à chaque fois qu'une nouvelle observation s'ajoute à la série.

Pour la révision des séries désaisonnalisées contemporaines, les organismes statistiques procèdent souvent de la façon suivante: ils maintiennent fixe l'estimation du facteur saisonnier contemporain depuis le moment où elle est établie jusqu'à la fin de l'année et révisent ensuite annuellement l'année courante et les années antérieures. Ainsi, les révisions de première année attribuables aux différences entre les filtres sont données par $R^{(0,0)}, R^{(1,0)}, \dots, R^{(11,0)}$, les révisions de seconde année par $R^{(12,0)}, R^{(13,1)}, \dots, R^{(23,11)}$, et ainsi de suite, où $R^{(k,k)}$ est définie par

$$R^{(k,k)} = [2\int_{1/2}^0 \|\Gamma^{(k)}(\omega) - \Gamma^{(k)}(\omega)\|^2 d\omega]^{1/2}, \quad \ell = 1, 2, \dots, n, k = 0, 1, 2, \dots, n - 12, \tag{17}$$

et $n = 42$ pour les filtres du X-11-ARMMI.

Le tableau 1 montre les révisions de première, de seconde et de troisième année apportées par le filtre contemporain du X-11-ARMMI (avec et sans extrapolations ARMMI), les extrapolations étant tirées d'un modèle ARMMI donné et de deux séries de paramètres (d'autres cas sont exposés dans Dagum (1987)). Le modèle ARMMI choisi est le modèle $(0,1,1) (0,1,1)_{12}$, c'est-à-dire $(1 - B) (1 - B^{12}) X'_t = (1 - \theta B) (1 - \theta B^{12}) a'_t$ où X'_t désigne la série initiale, B est l'opérateur de retard de sorte que $B^n X'_t = X'_{t-n}$, a'_t est un élément purement aléatoire qui représente les résidus, et θ et Θ désignent respectivement les paramètres non saisonnier et saisonnier.

Comme les plus fortes révisions par période se produisent à $\ell = 1$ et $\ell = 12$, on obtient-drait un meilleur modèle de révision en intégrant des révisions mensuelles et annuelles. Ainsi, (1) en soumettant les données de chaque mois (de janvier à novembre par exemple) à une désaisonnalisation contemporaine puis en révisant ces données une seule fois, au moment où les données du mois suivant sont connues, et (2) en appliquant une désaisonnalisation contemporaine aux données de décembre des qu'elles sont connues, puis en révisant les données de la première année et des années antérieures lorsque les données de janvier sont connues, nous devrions accroître la fiabilité du filtre appliqué dans l'année courante tout en conservant son homogénéité pour les comparaisons d'un mois à l'autre.

Les révisions de première année apportées par le filtre ayant fait l'objet d'une première révision mensuelle seraient donc $R^{(1,1)}, R^{(2,1)}, \dots, R^{(11,1)}$. La valeur de ces révisions figure dans le tableau 2 et, malgré une tendance très comparable à celle observée dans le tableau 1 (filtre appliqué aux facteurs saisonniers contemporains), les révisions du tableau 2 sont beaucoup moins marquées s'il n'y a pas d'extrapolation. Par ailleurs, l'amélioration est moins notable lorsqu'on utilise des extrapolations ARMMI. On n'a constaté aucune différence majeure dans le cas des révisions de seconde et de troisième année.

3.1 Estimation des variations liées aux jours ouvrables et modèles ARMMI dans le contexte de la désaisonnalisation contemporaine

Outre le choix du modèle de révision à appliquer, la désaisonnalisation contemporaine pose deux autres problèmes qui ont trait aux modèles ARMMI et aux variations liées aux jours ouvrables.

3. RÉVISION DES DONNÉES DÉSAISONNALISÉES CONTEMPORAINES

Statistique Canada a utilisé la désaisonnalisation contemporaine pour la première fois en 1975 dans le cadre de l'Enquête sur la population active. Progressivement, les organismes statistiques d'autres pays ont suivi l'exemple de Statistique Canada et ont adopté cette méthode. L'utilisation de facteurs saisonniers contemporains dans la désaisonnalisation courante nous amène à nous demander à quel rythme faut-il réviser les séries. Kenny et Durbin (1982) ont proposé d'effectuer les révisions après un mois et, par la suite, à chaque année civile. Dagum (1982c) a appuyé ces conclusions et a proposé en plus une révision additionnelle après six mois si la méthode de désaisonnalisation utilisée est la X-11-ARMMI sans extrapolations ARMMI.

Pour deux points quelconques au temps $t + k$, $t + \ell$ ($k < \ell$), les révisions des estimations de facteurs saisonniers et, partant, des valeurs désaisonnalisées sont définies par

(14)
$$r_{(\ell k)}^t = X_{(t)}^t - X_{(k)}^t, k < \ell.$$

Ces révisions rendent compte: (1) des innovations introduites par les nouvelles observations $X_{t+k+1}^t, X_{t+k+2}^t, \dots, X_{t+k+\ell}^t$ et (2) des différences entre les deux filtres de désaisonnalisation asymétriques $Y_{(t)}^{(k)}(B)$ et $Y_{(k)}^{(t)}(B)$. Si on pose $k = 0$ et qu'on fait varier ℓ de 1 à m , l'équation (14) produit une série de révisions apportées aux valeurs désaisonnalisées contemporaines pour des périodes ou des décalages différents. La *révision totale* de l'estimation du facteur saisonnier contemporain est donnée pour $\ell = m$. Si on pose $\ell = k + 1$ et qu'on fait varier k de 0 à $m - 1$, on obtient par l'équation (14) la série de *révisions apportées par période* à chaque valeur désaisonnalisée estimée avant qu'elle ne devienne finale. Si on pose $\ell = k + 12$ et qu'on fait varier k de 0 à $m - 12$, on obtient par l'équation (14) la série de révisions annuelles. Les révisions qui nous intéressent en l'occurrence sont celles causées par les différences entre les filtres, et il est possible d'analyser ces différences en examinant les fonctions de réponses en fréquence complexes des filtres correspondants. Comme dans l'équation (6), nous pouvons estimer les valeurs désaisonnalisées pour les années récentes au moyen du programme X-11-ARMMI (avec ou sans extrapolations ARMMI) par

(15)
$$X_{(n)}^t = \sum_{m}^{j=n} Y_{n,j} X_{t-j}^t = Y_{(n)}^{(t)}(B) X_t.$$

L'équation (15) représente un système linéaire où $X_{(n)}^t(n)$ est le produit de convolution de la série initiale X_t^t et d'une série de poids $Y_{n,j}$ appelée *la fonction de réponse à une impulsion* du filtre. On peut étudier les propriétés de cette réponse à l'aide de sa transformée de Fourier que l'on appelle *la fonction de réponse en fréquence complexe* et qui est définie par

(16)
$$\Gamma_{(n)}^{(n)}(\omega) = \sum_m^{j=-n} Y_{n,j} e^{-2\pi\omega j}, -1/2 \leq \omega \leq 1/2,$$

où ω est la fréquence exprimée en cycles par unité de temps. $\Gamma_{(n)}^{(n)}(\omega)$ décrit entièrement les effets du filtre linéaire sur la série initiale donnée. Dagum (1987) a calculé les révisions mensuelles et annuelles apportées par les filtres contemporains du X-11-ARMMI (avec ou sans extrapolations ARMMI), en se fondant sur la distance mathématique entre les diverses fonctions de réponse en fréquence complexes des filtres. Ces calculs nous permettent de constater une diminution rapide de l'importance des révisions mensuelles pour $\ell = 1, 2$, et 3, puis une diminution lente jusqu'à $\ell = 11$. On observe ensuite une forte augmentation à $\ell = 12$, puis une diminution à $\ell = 13$ et de nouveau une forte augmentation à $\ell = 24$, celle-ci suivie d'une diminution à $\ell = 25$. Dagum (1987) a montré que ces révisions mensuelles suivent la même tendance, qu'on utilise ou non les extrapolations ARMMI.

$$S'_{(t)} = \sum_{j=-2m}^j h_{tj} X_{t-j} = h^{(t)}(B) X_t, \quad m = 42, \quad (8)$$

où $h^{(t)}(B)$ désigne le filtre asymétrique saisonnier *prévu* et $\ell = 1, 2, \dots, 12$ pour une série mensuelle.

La révision d'une estimation d'un facteur saisonnier contemporain dépend de la distance entre le filtre contemporain et le filtre final, c'est-à-dire $d[h^{(0)}(B), h^{(m)}(B)]$, et des innovations des nouvelles observations $X_{t+1}, X_{t+2}, \dots, X_{t+m}$.

De même, la révision d'une estimation d'un facteur saisonnier prévu dépend de $d[h^{(t)}(B), h^{(m)}(B)]$ et des innovations introduites par $X_{t-\ell}, \dots, X_t, X_{t+1}, \dots, X_{t+m}$. Dagum (1982a et 1982b) a montré dans des études théoriques que

$$d[h^{(0)}(B), h^{(m)}(B)] < d[h^{(\ell)}(B), h^{(m)}(B)] \text{ for } \ell = 1, 2, \dots, 12. \quad (9)$$

La distance entre les deux filtres est définie comme l'écart quadratique moyen entre les fonctions de réponse en fréquence complexes des filtres pour l'ensemble des fréquences saisonnières. Une définition semblable est donnée dans la section suivante (équation 17), avec la racine de l'écart quadratique moyen.

L'inéquation (9) est vraie peu importe que l'on utilise des extrapolations ARMMI ou non. En outre, les deux études précitées ont montré que

$$d[h^{(0)}(B), h^{(m)}(B)] \text{ avec extrapolations ARMMI} < d[h^{(0)}(B), h^{(m)}(B)] \text{ sans extrapolations ARMMI} \quad (10)$$

de même,

$$d[h^{(0)}(B), h^{(m)}(B)] \text{ avec extrapolations ARMMI} < d[h^{(\ell)}(B), h^{(m)}(B)] \text{ sans extrapolations ARMMI} \quad (11)$$

pour $\ell = 1, 2, \dots, 12$.

Plusieurs études (Dagum (1978), Bayer et Wilcox (1981), Kenney et Durbin (1982), McKenzie (1984), Dagum et Morry (1984), Pierce (1980) et Pierce et McKenzie (1985)) ont montré que

$$r^{(0,m)} > r^{(\ell,m)}. \quad (12)$$

Mais dans quelques cas,

$$r^{(0,m)} > r^{(\ell,m)}. \quad (13)$$

L'inéquation (13) se vérifie lorsque les observations courantes de la dernière année sont fortement modifiées parce que X_t reçoit le poids le plus élevé dans les estimations de $S'_{(t)}$. Du point de vue de la révision totale des estimations de facteurs saisonniers, nous pouvons classer les quatre modes de désaisonnalisation en nous fondant sur les résultats des études empiriques mentionnées ci-dessus: le mode (i) (facteurs saisonniers contemporains avec extrapolations ARMMI) produit la révision totale la plus faible; le mode (iii) (facteurs saisonniers contemporains sans extrapolations ARMMI) vient au second rang; le mode (ii) (facteurs saisonniers prévus avec extrapolations ARMMI) vient au troisième rang, suivi du mode (iv) (facteurs saisonniers prévus sans extrapolations ARMMI).

Bien que les organismes statistiques se servent des quatre modes pour obtenir des valeurs désaisonnalisées courantes, la fréquence d'application de chacun de ces modes varie selon les organismes. Par exemple, Statistique Canada utilise surtout le mode (i), puis le mode (iii) tandis que le U.S. Bureau of Labor utilise surtout le mode (ii), puis le mode (iv). Les quatre modes ne produisent pas tous la même valeur désaisonnalisée courante et ne renferment pas non plus le même degré d'erreur.

Suivant l'hypothèse d'un modèle de décomposition additif, on désaisonnalisera une valeur courante X_t par l'équation

(1)
$$X_t^{(0)} = X_t - S_t^{(0)},$$
 où $S_t^{(0)}$ désigne l'estimation d'un facteur saisonnier prévu; ou

(2)
$$X_t^{(0)} = X_t - S_t^{(0)},$$
 où $S_t^{(0)}$ désigne l'estimation d'un facteur saisonnier contemporain.

La valeur courante désaisonnalisée deviendra "finale", c'est-à-dire qu'elle ne sera plus révisée, lorsque m observations auront été ajoutées à la série. Ainsi,

(3)
$$X_t^{(m)} = X_t - S_t^{(m)},$$
 où $S_t^{(m)}$ désigne l'estimation finale d'un facteur saisonnier.

On peut donc exprimer la révision totale d'un facteur saisonnier contemporain et d'un facteur saisonnier prévu par les formules suivantes:

(4)
$$r_{(0,m)}^t = S_t^{(0)} - S_t^{(m)}, \quad m > 0;$$

(5)
$$r_{(l,m)}^t = S_t^{(l)} - S_t^{(m)}, \quad m > 0 > l$$

Suivant l'hypothèse d'un modèle de décomposition additif et du non-remplacement des valeurs extrêmes, $S_t^{(m)}$, la valeur finale d'un facteur saisonnier tirée d'une série $X_{t-m}, \dots, X_t, \dots$, X_{t+m} , peut être défini comme étant

(6)
$$S_t^{(m)} = \sum_{j=-m}^m h_{m,j} X_{t-j} = h^{(m)}(B) X_t,$$

où $h_{m,j} = h_{m,-j}$ sont les poids des moyennes mobiles symétriques qui doivent être appliquées à la série. Le terme $h^{(m)}(B)$ désigne le filtre linéaire correspondant qui utilise l'opérateur de retard B de telle sorte que $B^n = X_{t-n}$. Young (1968) a montré que l'étendue de ce filtre symétrique est de 145 poids pour les séries mensuelles mais que l'on peut en obtenir une bonne approximation au moyen de 85 poids parce que la valeur des poids attribués aux observations éloignées est très faible; par conséquent, $m = 42$.

Étant donné l'équation (6), nous pouvons exprimer l'estimation d'un facteur saisonnier contemporain $S_t^{(0)}$ et l'estimation d'un facteur saisonnier prévu $S_t^{(l)}$ par

(7)
$$S_t^{(0)} = \sum_{j=-2m}^0 h_{0,j} X_{t-j} = h^{(0)}(B) X_t, \quad m = 42,$$

où $h^{(0)}(B)$ désigne le filtre asymétrique saisonnier contemporain; et

Problèmes courants sur la désaisonnalisation

ESTELA BEE DAGUM¹

RÉSUMÉ

Dans cet article, l'auteur analyse trois questions qui, depuis une dizaine d'années, intéressent vivement les spécialistes de la désaisonnalisation qui oeuvrent dans les organismes statistiques. Ces trois questions sont: (1) le choix entre facteurs saisonniers contemporains et facteurs saisonniers prévus pour la désaisonnalisation courante; (2) la définition d'un modèle de révisions optimal pour les séries désaisonnalisées au moyen de facteurs saisonniers contemporains; et (3) le lissage de données désaisonnalisées très irrégulières.

MOTS CLÉS: Facteurs saisonniers contemporains et facteurs saisonniers prévus; révisions; filtres de la tendance-cycle; lissage.

1. INTRODUCTION

Dans les dix dernières années, trois questions importantes touchant la désaisonnalisation ont retenu l'attention des responsables des organismes statistiques: (1) la désaisonnalisation de valeurs courantes; (2) la révision de données désaisonnalisées contemporaines; et (3) le lissage de séries désaisonnalisées très irrégulières.

Cet article vise principalement à analyser chacune de ces questions en fonction de la méthode de désaisonnalisation X-11-ARMMI élaborée par Dagum (1980) et appliquée par Statistique Canada et d'autres organismes statistiques à l'étranger.

Dans la section 2, nous analysons les quatre modes d'application du programme X-11-ARMMI qui vise à produire des valeurs désaisonnalisées courantes. Dans la section 3, nous allons surtout examiner les révisions des données désaisonnalisées contemporaines, qui découlent de l'application des filtres linéaires de la méthode X-11-ARMMI. Enfin, dans la section 4, nous examinons les caractéristiques des filtres de lissage (tendance-cycle) de la X-11-ARMMI.

2. DÉSAISONNALISATION DE VALEURS COURANTES

Une valeur courante peut être désaisonnalisée à l'aide d'un facteur saisonnier contemporain ou d'un facteur saisonnier prévu.

On obtient l'estimation d'un facteur saisonnier contemporain (facteur ou effet qui varie selon qu'il s'agit respectivement, par hypothèse, d'un modèle multiplicatif ou additif) en désaisonnalisant, chaque fois qu'une nouvelle observation est enregistrée, toutes les données antérieures, y compris l'observation en question. Par ailleurs, on détermine un facteur saisonnier prévu au moyen d'une série qui se termine dans l'année précédente. D'une manière générale, on projette par exemple des facteurs saisonniers pour l'année $T + 1$ au moyen d'une série de données qui se termine en décembre de l'année précédente T .

Il existe quatre modes d'application du programme X-11-ARMMI, qui permettent de produire une valeur désaisonnalisée courante (dernière observation). Ce sont: (i) les extrapolations ARMMI avec facteurs saisonniers contemporains; (ii) les extrapolations ARMMI avec facteurs saisonniers prévus; (iii) les facteurs saisonniers contemporains sans extrapolations ARMMI; et (iv) les facteurs saisonniers prévus sans extrapolations ARMMI.

¹ Estela Bee Dagum, Division des séries chronologiques recherche et analyse, Direction de la méthodologie, Statistique Canada, 13^e étage, Immeuble R.H. Coats, Ottawa (Ontario), Canada, K1A 0T6.

Par ailleurs, les estimateurs de variance pour les groupes aléatoires sont essentiellement des estimateurs non biaisés de la variance de l'estimation du rendement et de la production. Cependant, ces estimateurs sont beaucoup moins stables que $V(\bar{y}_r)$ et $V(\bar{X}_r)$. C'est pour-quoi ces derniers sont préférés aux estimateurs pour les groupes aléatoires.

L'enquête énumérative de juin constitue la première phase de l'enquête objective sur le rendement. Les méthodes d'échantillonnage utilisées dans cette enquête sont simples et, comme en fait foi la simulation de Monte Carlo, produisent des estimations de superficie justes. Par conséquent, il ne convient pas d'apporter des modifications à la première phase de l'enquête objective sur le rendement.

On peut toutefois envisager un certain nombre de modifications pour la seconde phase de cette enquête. À l'heure actuelle, on estime le rendement au moyen d'un sondage à deux phases, dans lequel est utilisé un estimateur par quotient combiné. Pour les États où l'échan-tilion est relativement grand, il conviendrait d'envisager un échantillonnage indépendant à la seconde phase pour chaque strate ou des groupes de strates, ainsi que l'utilisation d'un estimateur par quotient distinct.

Pour obtenir des estimateurs non biaisés de la variance, il conviendrait de remplacer l'échan-tillonnage systématique effectué dans la seconde phase. À l'heure actuelle, l'échantillonnage des segments à la seconde phase pour l'estimation du rendement se fait par ordinateur et vise tous les États américains. Il serait donc relativement facile d'adopter une méthode d'échan-tillonnage qui utiliserait des probabilités conjointes connues. Des estimateurs semblables à ceux qui sont recommandés pour le plan de sondage qui nous intéresse pourraient toujours convenir si les probabilités de sélection demeuraient les mêmes. Fuller (1970) décrit une méthode d'échantillonnage qui peut être informatisée et qui comprend le calcul de probabi-lités conjointes de sélection; en outre, cette méthode utilise des probabilités de sélection détec-minées et assure un degré de contrôle comparable à celui qu'offre l'échantillonnage systématique.

REMERCIEMENTS

Cette étude a été réalisée en partie grâce à un contrat de recherche coopérative (n° 58-319T-1-0054X) passé avec le National Agricultural Statistics Service du Département de l'agri-culture des États-Unis. Nous tenons à exprimer notre reconnaissance aux arbitres pour leurs commentaires utiles.

BIBLIOGRAPHIE

PECOSO, R. (1978). Cluster analysis as an aid in creating paper strata. Statistical Reporting Service, U.S. Department of Agriculture.

PECOSO, R., et JOHNSON, V. (1981). The new California area frame: A statistical study. Statistical Reporting Service, U.S. Department of Agriculture.

FULLER, W. A. (1970). Sampling with random stratum boundaries. *Journal of the Royal Statistical Society, Sér. B*, 32, 209-226.

HOUSEMAN, E. E. (1975). Area frame sampling in agriculture. Statistical Reporting Service, U.S. Department of Agriculture.

PRATT, W. L. (1984). The use of interpenetrating sampling in area frames. Statistical Reporting Ser-vice, U.S. Department of Agriculture.

WOLTER, K. M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

pour le j -ième échantillon de rendement ($j = 1, 2$) dans le i -ième échantillon de l'enquête énumérative de juin ($i = 1, \dots, 500$). L'erreur type estimée de l'écart était de 0,58. Ainsi, la valeur moyenne de $V(\bar{y}_j)$ se situe à moins de 1,5 erreurs types de la variance estimée de \bar{y}_j . Par ailleurs, la valeur moyenne de l'estimation de la variance de \bar{y}_j se situe à moins de 2% de la variance observée dans la simulation de Monte Carlo.

Les estimateurs de la variance de \bar{y} par les groupes aléatoires avaient un biais peu significatif. Dans la simulation de Monte Carlo, les moyennes des estimateurs $V_{gs}(\bar{y})$ et $V_{g10}(\bar{y})$ étaient respectivement de 9 et de 7% supérieures aux variances observées. Ces écarts ne sont pas significativement différents de zéro et sont comparables à ceux obtenus pour l'estimateur $V(\bar{y}_j)$. Celui-ci, toutefois, est un estimateur de variance beaucoup plus stable, le coefficient de variation pour cet estimateur étant d'environ 30%, comparativement à 75% pour $V_{gs}(\bar{y})$. Comme prévu (Wolter 1985), un accroissement du nombre de groupes aléatoires a amené une diminution du coefficient de variation de l'estimateur de la variance par les groupes aléatoires. En effet, le coefficient de variation pour $V_{g10}(\bar{y})$ était de 50%. Les variations entre groupes aléatoires et entre échantillons de rendement de l'enquête énumérative de juin ont expliqué en majeure partie la variance des estimateurs de variances par les groupes aléatoires.

4. CONCLUSIONS

Les analyses montrent que les estimateurs du rendement moyen et de la production totale d'un État, qui sont actuellement utilisés par le National Agricultural Statistical Service sont satisfaisants. En revanche, nous avons vu que les estimateurs de variance simples $V_2(\bar{y})$ et $V^*(\bar{y})$ étaient entachés d'un biais négatif dont l'importance dépendait de la variance à l'intérieur des segments de sondage à l'intérieur des segments. L'estimateur $V_2(\bar{y})$ a produit une estimation qui était près de 40% inférieure à la vraie variance de \bar{y} , tandis que l'estimateur $V^*(\bar{y})$ en a produit une qui était de 18% inférieure à la vraie variance de \bar{y} pour la population simulée.

On a élaboré les estimateurs \bar{y}_j et \bar{y}_j pour l'estimation de rendement d'un échantillonage à deux phases, selon lequel les segments qui sont identifiés comme des segments de culture du coton dans la première phase (enquête énumérative de juin) sont sous-échantillonnés dans la seconde phase afin d'estimer le rendement. On estime la probabilité non conditionnelle qu'un segment soit prélevé en vue de recevoir une unité secondaire à l'intérieur d'une strate, π_{hk}^* , en supposant que cette probabilité est proportionnelle à la probabilité conditionnelle de sélection de segments dans la seconde phase de l'échantillonnage. Suivant cette hypothèse, on a défini l'estimateur par quotient combiné du rendement moyen avec probabilités inégales, \bar{y}_j , et l'estimateur de sa variance, $V(\bar{y}_j)$. L'estimateur de l'agrégat \bar{y}_j est un estimateur par produit de la production moyenne d'un segment pour un sondage à deux phases, où l'estimateur de la moyenne de la variable auxiliaire (superficie consacrée à la culture étudiée) est tiré de l'enquête énumérative de juin (première phase du sondage). L'estimateur de variance $V(\bar{y}_j)$ est un estimateur de la variance de \bar{y}_j pour un échantillonnage double stratifié (sondage à deux phases).

La simulation de Monte Carlo a montré que \bar{y}_j et \bar{y}_j produisaient des estimations comparables à celles découlant des estimateurs courants \bar{y} et \bar{y} , $V(\bar{y}_j)$ et $V(\bar{y}_j)$ sont tous deux des estimateurs de variance précis pour des échantillons de la taille de ceux qu'utilise normalement le NASS. Ces résultats sont attribuables en partie à l'exactitude avec laquelle les superficies cultivées moyennes sont estimées au moyen de l'enquête énumérative de juin. Des estimations de superficie justes produisent des estimations de sélection qui se rapprochent des probabilités de sélection non conditionnelles. En outre, le fait que l'estimateur soit sous forme de quotient amenuise les effets de la substitution d'estimateurs aux probabilités non conditionnelles réelles.

Tableau 6

Propriétés des estimations du rendement à l'acre et des estimations de variances selon la simulation de Monte Carlo¹

Estimateur					
\bar{y}	$V_2(\bar{y})$	$V_{g5}(\bar{y})$	$V_{g10}(\bar{y})$	\bar{y}_r	$V(\bar{y}_r)$
Moyenne	79.74	7.21	12.62	12.39	79.76
Variance totale	11.57	0.99	74.58	36.86	11.56
Variance entre les EBJ	7.60	0.48	6.10	4.56	7.64
Variance à l'intérieur					
des EBJ	3.97	0.51	68.48	32.30	3.92
					4.90

¹ Deux échantillons de l'enquête objective sur le rendement ont été simulés pour chacun des 500 échantillons de l'enquête énumérative de juin.

Tableau 7

Propriétés des estimations de la production et des estimations de variances selon la simulation de Monte Carlo¹

Estimateur ²					
\bar{y}	$V^*(\bar{y})$	$V_{g5}(\bar{y})$	$V_{g10}(\bar{y})$	\bar{y}_r	$V(\bar{y}_r)$
Moyenne	73.04	40.85	48.99	48.53	73.07
Variance totale	49.69	82.52	1245.10	608.80	49.58
Variance entre les EBJ	46.35	78.17	50.82	208.48	46.30
Variance à l'intérieur					
des EBJ	3.34	4.35	1194.28	400.32	3.28
					23.38

¹ Deux échantillons de l'enquête objective sur le rendement ont été simulés pour chacun des 500 échantillons de l'enquête énumérative de juin. Il y avait $N = 92,240$ segments dans la population simulée.
² L'estimateur \bar{y} est exprimé en millions d'unités et les variances le sont dans les unités correspondantes.

Par ailleurs, cet écart est dû au fait que $V^*(\bar{y})$ ne tient pas compte de la covariance de \bar{M}_n et de \bar{y} . Dans l'exemple qui nous occupe, le biais lié au second facteur contrebalance en partie celui lié au premier.

L'utilisation de l'expression (3.16), $V(\bar{y}_r)$, pour estimer la variance de \bar{y}_r et de l'expression (3.21), $V(\bar{y}_r)$, pour estimer la variance de \bar{y} , a donné des résultats beaucoup plus satisfaisants que ceux obtenus avec les estimateurs présentement en usage. La moyenne des estimations $V(\bar{y}_r)$ a été de 12.51 dans la simulation de Monte Carlo, ce qui représente une surestimation d'environ 7% par rapport à la variance observée de \bar{y}_r (11.57). Environ le tiers de cet écart (l'équivalent de 2 à 4 pour cent) peut s'expliquer par l'utilisation d'un échantillonage sans remise aux deux premiers degrés de l'échantillonnage. Le reste (l'équivalent de 4 pour cent environ) est faible par rapport à l'erreur type de l'écart estimé. On a estimé la variance de l'écart en estimant la variance de la moyenne de z_{ij} , où

$$z_{ij} = (\bar{y}_{n,r(ij)} - 79.76)^2 - V(\bar{y}_{n,r(ij)}), \quad (4.10)$$

De même, posons $Y^{(\alpha)}$ comme l'estimateur de la production totale pour l' α -ième groupe aléatoire:

$$Y^{(\alpha)} = N M_{n^{(\alpha)}} Y^{(\alpha)}, \quad (4.8)$$

ou

$$\bar{M}_{n^{(\alpha)}} = \sum_{L=1}^h W_h K_{h^{(\alpha)}}^{-1} \sum_{K_{h^{(\alpha)}}=1}^K M_{h^{(\alpha)}}^{K_{h^{(\alpha)}}},$$

$M_{h^{(\alpha)}}^{K_{h^{(\alpha)}}}$ est la superficie (en acres) consacrée à la culture du coton dans le segment k de la strate h pour le groupe aléatoire α et $K_{h^{(\alpha)}}$ est le nombre de segments compris dans la strate h pour l' α -ième groupe. L'estimateur de la variance de Y par les groupes aléatoires est alors défini comme étant

$$V_{gy}(Y) = \gamma(\gamma - 1)^{-1} \sum_{\gamma}^{\alpha=1} (Y^{(\alpha)} - \bar{Y})^2. \quad (4.9)$$

Les tableaux 6 et 7 présentent un sommaire des résultats de la simulation de Monte Carlo pour les estimateurs du rendement et de la production. On y trouve les valeurs moyennes des estimations et leurs estimations de variances pour les 1,000 échantillons simulés de l'enquête objective sur le rendement. En simulant deux échantillons de l'enquête objective sur le rendement pour chaque échantillon simulé de l'enquête énumérative de juin, nous avons pu estimer la variance entre enquêtes énumératives et la variance à l'intérieur des enquêtes énumératives. L'estimateur actuellement en usage, \bar{y} , défini en (3.1), et l'estimateur par quotient combiné \bar{y}_r , défini en (3.17), qui est fondé sur les probabilités π_{hk}^* établie à partir des résultats de l'enquête énumérative de juin, produisent tous deux des estimations de précision comparable (voir tableau 6). Cette similitude est en partie attribuable à l'exactitude avec laquelle on estime les probabilités de sélection non conditionnelles dans chaque échantillon.

Comme nous l'avons vu dans la section 3.2, l'estimateur de la variance conditionnelle $V_2(\bar{y})$ produit une sous-estimation de $V(\bar{y})$. Dans cette étude de simulation $V_2(\bar{y})$ a produit une estimation qui est de 38% inférieure à la variance observée de \bar{y} . En effet, la variance observée de \bar{y} était de 11.57, comparativement à une moyenne de 7.21 pour $V_2(\bar{y})$. Cette sous-estimation de la variance a été constatée dans l'ensemble des échantillons. La variance estimée de $V_2(\bar{y})$ était de 0.99, pour des valeurs de $V_2(\bar{y})$ allant de 3.85 à 11.24 pour les 1,000 observations. Ainsi, la valeur maximale observée pour l'estimation de la variance conditionnelle était inférieure à la vraie variance.

Suivant l'hypothèse d'un échantillonnage de segments avec probabilité proportionnelle à la taille et remise en seconde phase, $V_2(\bar{y})$ est un estimateur sans biais de la variance conditionnelle de \bar{y} , étant donné l'échantillon de segments prélevés au premier degré, comme nous l'avons vu dans la section 3.2. Selon la simulation de Monte Carlo, une estimation de l'espérance mathématique de la variance conditionnelle de \bar{y} , $V_2(\bar{y})$, est 3.97. L'écart appréciable entre cette estimation et la moyenne indiquée au tableau 6 (3.97 contre 7.21) peut être attribuable au fait que l'estimateur $V_2(\bar{y})$ ne tient pas compte des effets de la stratification dans la population (voir tableaux 2 et 3) et que $V_2(\bar{y})$ est calculée suivant l'hypothèse d'un échantillonnage de segments avec remise au second degré du sondage.

L'estimateur (3.9), $V^*(\bar{Y})$, produit une sous-estimation de la variance non conditionnelle de \bar{Y} . En effet, la variance observée de \bar{Y} dans la simulation de Monte Carlo est 49.69 (millions)² tandis que la valeur moyenne de $V^*(\bar{Y})$ n'est que 40.85 (millions)². Cette sous-estimation (18%) de la vraie variance est attribuable à un certain nombre de facteurs. Nous avons vu plus tôt que $V_2(\bar{y})$ comporte un biais négatif lorsqu'il sert d'estimateur de $V(\bar{y})$.

Tableau 5
 Estimations de la superficie ensemencée de coton établies à partir de 500 échantillons simulés de l'enquête énumérative de juin

$V(\bar{A}_n)$	Moyenne	Intervalle	Variance
	9.93	8.13 - 12.21	0.66
	0.64		0.016

La superficie moyenne ensemencée de coton (par segment) pour la population simulée est de 9.94 acres tandis que la moyenne des estimations des 500 échantillons était de 9.93 acres. La variance réelle de l'estimateur stratifié \bar{A}_n est 0.63 tandis que la variance moyenne estimée pour les 500 échantillons simulés était 0.64. A cause de la faible variabilité de l'estimation de la superficie consacrée à la culture du coton, π_{hk}^* produit une estimation stable de la probabilité non conditionnelle que le segment k de la strate h soit échantillonné pour recevoir au moins une unité secondaire.

Outre les estimateurs analysés ci-dessus, nous avons construit des estimateurs de la variance par les groupes aléatoires. Deux ensembles de groupes aléatoires ont été formés pour chaque échantillon de l'enquête objective sur le rendement. Le premier ensemble contenait cinq groupes ($\gamma = 5$) et l'autre, dix ($\gamma = 10$). Les groupes aléatoires ont été créés par la formation de sous-ensembles avec les unités primaires d'échantillonnage (c'est-à-dire les segments) à l'intérieur de chaque strate d'exploitation. On a formé le premier groupe de chaque ensemble en tirant un échantillon aléatoire simple sans remise de taille $K_{h(\gamma)} = n_h/\gamma$ dans chaque strate ($h = 1, \dots, 28$) de l'échantillon parent de l'enquête énumérative de juin. Le second groupe a été formé de la même manière par tirage de $K_{h(\gamma)}$ segments parmi les $n_h - K_{h(\gamma)}$ segments qui restaient dans chaque strate. Les autres groupes aléatoires ont été formés d'une manière semblable. Comme l'échantillon de la strate d'exploitation n° 3107 ne comptait que cinq segments ($n_h = 5$), on a répété les valeurs de superficie et de rendement des cinq segments pour obtenir les dix observations nécessaires à la formation des dix groupes quand $\alpha = 10$. Soit D_α le nombre d'unités secondaires ensemencées de coton qui ont été prélevées dans le groupe aléatoire α ($\alpha = 1, \dots, \gamma$) au cours de l'enquête objective sur le rendement. Soit $y^{(\alpha)}$ l'estimateur de rendement obtenu pour l' α -ième groupe aléatoire:

$$(4.6) \qquad y^{(\alpha)} = D_\alpha^{-1} \sum_{i=1}^I Y_{t(\alpha)}^i,$$

où $y^{(\alpha)}$ est l'équivalent de l'équation (3.3) pour l' α -ième groupe. L'estimateur de la variance de y par les groupes aléatoires prend la forme suivante:

$$(4.7) \qquad V_{gy}(y) = \gamma(\gamma - 1)^{-1} \sum_{\gamma}^{ \alpha = 1} (y^{(\alpha)} - y)^2.$$

Cet estimateur est légèrement biaisé dans le cas de l'estimateur de rendement se rapportant à la série de dix groupes parce que la strate n° 3107 ne comporte que cinq observations et que celles-ci se répètent dans les divers groupes.

Tableau 4

Analyse de variance pour les données de l'enquête objective sur le rendement de 1983

Source	Degrés de Sommes			Composante	Pourcentage
	liberté	des carrés	moyen		
Strate	26	80,193	3,084.3	187.3	14
Segments à l'intérieur de la strate	85	124,086	1,459.8	378.0	28
Résidu	103	79,991	776.6	776.6	58
Total	214	284,270		1,341.9	100

Cinq cents échantillons de l'enquête énumérative de juin ont été tirés de la population simulée au moyen d'un échantillonnage aléatoire stratifié. Deux cent soixante-quinze segments ont été échantillonnés dans chaque cas. Le nombre de segments prélevés dans chaque strate était identique au nombre prélevé pour l'enquête énumérative de juin 1983 (voir tableau 2). Pour chacun des échantillons simulés, on a calculé la superficie moyenne (en acres) des segments dans la population et les probabilités conditionnelles, $\{\pi_{hk}\}$, de (3.12), que les segments de l'échantillon recevaient des parcelles par suite d'un tirage. Les probabilités conditionnelles ainsi calculées ont été utilisées au second degré de l'échantillonnage systématique avec un seul départ et probabilité proportionnelle à la taille estimée (voir section 2). Afin de simuler des échantillons de l'enquête objective de rendement on a prélevé 220 unités secondaires en se servant de cette méthode d'échantillonnage systématique. Les unités secondaires ainsi prélevées constituent en soi un échantillon de l'enquête objective sur le rendement. Deux échantillons de ce genre ont été simulés pour chacun des 500 échantillons de la population simulée de l'enquête énumérative de juin.

Lorsqu'un segment était sélectionné pour recevoir une unité secondaire, on simulait le rendement (nombre de plants aux 100 pi²) observé à l'intérieur d'un champ en supposant l'invariabilité du coefficient de variation à l'intérieur de chaque segment. Le nombre de plants observé était défini comme étant

$$y_{hkt} = \bar{c}_{hk} + s_w y_N^{-1} \bar{c}_{hk} f_{hkt}, \tag{4.3}$$

où y_{hkt} est le nombre moyen estimé de plants aux 100 pi² pour la k -ième unité secondaire du segment k de la strate h et f_{hkt} est distribuée suivant une loi normale $N(0, 1)$. L'erreur type à l'intérieur des segments est la racine carrée de $s_w^2 = 776.6$, qui figure dans le tableau 4, et \bar{y}_N est le nombre moyen global de plants par parcelle. Si y_{hkt} équivalait à moins de 10% de la moyenne pour la strate, on fixait sa valeur à $(.10)\bar{c}_{hk}$. De même, si y_{hkt} équivalait à plus de 190% de la moyenne pour la strate, sa valeur était fixée à $(1.9)\bar{c}_{hk}$.

Le tableau 5 donne un résumé des résultats des simulations pour les superficies ensemencées de coton. La superficie moyenne estimée (en acres) par segment est

$$\bar{A}_n = \sum_L^h W_h n_h^{-1} \sum_n^k A_{hk}, \tag{4.4}$$

et la variance estimée correspondante est

$$V(\bar{A}_n) = \sum_L^h W_h^2 n_h^{-1} (n_h - 1)^{-1} \sum_n^k (A_{hk} - \bar{A}_h)^2. \tag{4.5}$$

Tableau 3
Nombre moyen de plants aux 100 pi² selon l'enquête objective
sur le rendement de 1983 pour le coton en Californie
et dans la population simulée

Strate	Nombre moyen de plants aux 100 pi ²	Enquête objective sur le rendement de 1983	Population simulée
--------	---------------------------------------------------	-----------------------------------------------	-----------------------

1314	78	76
1315	80	80
1316	67	68
1317	72	73
1318	80	80
1319	93	93
1320	92	91
1321	70	69
1322	84	84
1323	72	71
1713	118	117
1714	96 ¹	95
1715	96 ¹	93
1716	96 ¹	86
1717	96 ¹	96
1718	139	140
1719	96 ¹	97
1720	96 ¹	97
1721	89	86
1722	79	79
1723	84	85
1906	98	98
1907	67	67
1908	53	53
2010	118	118
2011	47	47
3107	80 ²	79
4110	60	59

¹ Moins de trois unités secondaires observées; ce chiffre représente la moyenne pour les unités secondaires de la strate d'exploitation 17.
² Moins de trois unités secondaires observées; ce chiffre représente la moyenne pour les unités secondaires de toutes les strates.

composante de la variance liée à la strate est tenue pour constante, 67% de la variation à l'intérieur des segments est attribuable à la variabilité entre les unités secondaires. Lorsqu'il avait été établi qu'un segment comprenait des champs de coton, le nombre moyen de plants aux 100 pi² pour le segment *hk* était simulé par

$$\bar{c}_{hk} = \bar{c}_h + e_{hk},$$

(4.2)

où \bar{c}_h est le nombre moyen de plants aux 100 pi² pour la strate *h*, e_{hk} est distribuée suivant une loi normale $N(0, s_h^2)$, et $s_h^2 = 378.0$. Si la moyenne obtenue par la simulation (\bar{c}_{hk}) équivaut à moins de 10% de la moyenne pour la strate, on fixait la valeur de \bar{c}_{hk} à $(.10)\bar{c}_h$. Le tableau 3 permet de comparer les moyennes obtenues pour la population simulée avec celles obtenues par l'enquête objective sur le rendement de 1983. La moyenne globale pour la population simulée était $\bar{y}_N = 79.6$.

Tableau 2
Estimations de la superficie ensemencée de coton selon l'enquête énumérative de juin 1983 en Californie et la superficie ensemencée de coton pour la population simulée

Strate	Taille du segment visée (acres)	Nombre de segments dans la strate	Nombre de segments échant. en 1983	1983		1983	
				Proportion de segments ensemencés de coton	Population simulée	Superficie moyenne ensemencée de coton dans les segments ensemencés de coton	Population simulée
1314	640	291	10	60	197	200	200
1315	640	291	10	100	354	348	348
1316	640	291	10	90	89	173	173
1317	640	291	10	90	149	148	148
1318	640	291	10	50	481	422	422
1319	640	291	10	20	249 ¹	260	260
1320	640	291	10	90	154	155	155
1321	640	291	10	60	270	274	274
1322	640	291	10	70	71	210	210
1323	640	291	10	80	79	279	279
1713	320	432	10	30	125	122	122
1714	320	432	10	30	31	57	57
1715	320	432	10	20	86 ²	84	84
1716	320	432	10	10	8	89	89
1717	320	432	10	40	38	27	27
1718	320	432	10	30	29	144	144
1719	320	432	10	30	31	67	67
1720	320	432	10	30	30	35	35
1721	320	432	10	30	29	138	138
1722	320	432	10	50	47	131	131
1723	320	432	10	40	76	76	76
1906	640	362	10	70	73	127	127
1907	640	362	10	70	74	194	194
1908	640	362	10	80	83	246	246
2010	640	649	10	30	31	306	306
2011	640	649	10	40	41	165	165
3107	160	1,847	5	20	22	25	25
4110	2,560	1,044	10	10	178	165	165

¹ Moins de trois segments échantillonnés; ce chiffre représente la moyenne pour tous les segments des sous-strates de la strate d'exploitation 13.
² Moins de trois segments échantillonnés; ce chiffre représente la moyenne pour tous les segments des sous-strates de la strate d'exploitation 17.
³ Moins de trois segments échantillonnés; ce chiffre représente la superficie approximative ensemencée de coton pour cette strate agro-urbaine.

ment étant difficiles à obtenir, on s'est servi d'une variable qui est une composante importante de ces estimations, soit le nombre de plants aux 100 pi². Dans l'enquête objective sur le rendement de 1983, le nombre moyen estimé de plants aux 100 pi² pour l'ensemble de la population était de 79.6. Le tableau 3 donne, pour chaque strate étudiée, le nombre moyen de plants aux 100 pi² selon l'enquête de 1983. La moyenne pour chaque strate est fondée sur toutes les unités secondaires de la strate qui ont été prélevées par échantillonnage avec probabilité proportionnelle à la taille estimée.

Une analyse de variance portant sur les données de 1983 (tableau 4) a montré que 28% de la variabilité totale entre les unités secondaires était attribuable à des différences entre segments à l'intérieur des strates ($s_b^2 = 378.0$), tandis que 58% de cette variabilité était attribuable à la variabilité entre unités secondaires à l'intérieur des segments ($s_w^2 = 776.6$). Si la

où $V\{M_n\}$ est l'estimateur de la variance d'une moyenne stratifiée. L'équation (3.21) définit un estimateur de la variance de l'estimation de la production totale d'un Etat pour un échantillonnage double stratifié. Contrairement à l'estimateur $V^*(Y)$ défini en (3.9), l'estimateur (3.21) ne repose pas sur l'hypothèse de l'absence de corrélation entre l'estimateur du rendement et l'estimateur de la superficie. L'équation (3.21) comporte également un estimateur non conditionnel de la variance du rendement.

3.4. Comparaison d'estimateurs par la méthode de Monte Carlo

Nous avons procédé à une étude par la méthode de Monte Carlo pour illustrer les différences entre divers estimateurs. À cette fin, nous nous sommes servis des données sur la superficie ensemencée de coton tirées de l'enquête énumérative de juin 1983 en Californie et des données de l'enquête objective sur le rendement correspondant de 1983. Pour les besoins de notre analyse, nous avons considéré que le coton était cultivé dans 28 strates. Le tableau 2 donne la répartition de cette culture entre les 28 strates, conformément aux données de l'enquête énumérative de 1983. Fecso et Johnson (1981) ont décrit les six modes d'exploitation du sol, qui sont identifiés par les deux premiers chiffres du numéro de la strate; ces modes d'exploitation sont définis ci-dessous avec le code correspondant:

- 1300 – au moins 50 pour cent des terres sont cultivées; principalement des cultures générales et au plus 10 pour cent de la superficie consacrée à la culture des fruits ou des légumes;
- 1700 – au moins 50 pour cent des terres sont cultivées; principalement culture des fruits, des noix ou du raisin combinée à des cultures générales;
- 1900 – au moins 50 pour cent des terres sont cultivées; principalement culture des légumes combinée à des cultures générales;
- 2000 – de 15 à 50 pour cent des terres sont cultivées; culture extensive et foin;
- 3100 – zones résidentielles et terres agricoles; plus de 20 logements au mille carré;
- 4100 – moins de 15 pour cent des terres sont cultivées; principalement de grands pâturages privés.

Nous avons simulé une population à partir des résultats de l'enquête énumérative de juin 1983. Le tableau 2 permet de comparer les caractéristiques de la population simulée aux résultats de l'enquête. Dans la population simulée, on pouvait dire que du coton était cultivé dans le segment k ($k = 1, \dots, N_h$) de la strate h ($h = 1, \dots, 28$) si $X_{hk} = 1$, où X_{hk} est une variable (aléatoire) de Bernoulli (p_h) indépendante, p_h étant la proportion de segments de la strate h où, selon l'enquête énumérative de 1983, du coton était cultivé.

La seconde étape de la création de la population consistait à attribuer une superficie (en acres) ensemencée de coton aux segments pour lesquels $X_{hk} = 1$. On a donc calculé une série de ratios mettant en relation la superficie ensemencée de coton dans un segment et la superficie moyenne des segments pour les strates d'exploitation qui renfermaient plus d'un segment ensemencées de coton selon l'enquête énumérative de juin 1983. Cet ensemble de ratios a servi à déterminer la superficie ensemencée de coton dans les segments où du coton était cultivé. Si X_{hk} était égal à 1, un ratio (r_{hk}) était prélevé dans l'ensemble de ratios observés de telle sorte que la probabilité de sélection était la même pour tous les ratios de l'ensemble. La superficie (en acres) ensemencée de coton dans le segment hk , M_{hk} , était définie par

$$M_{hk} = r_{hk} \bar{M}_h, \tag{4.1}$$

où \bar{M}_h était la superficie moyenne consacrée à la culture du coton pour les segments de la strate h où, selon l'enquête énumérative de juin 1983, du coton était cultivé. (Voir tableau 2.) Les résultats de l'enquête objective sur le rendement de 1983 pour le coton ont servi à simuler les observations de rendement à l'intérieur des segments. Les estimations de rende-

$Y_{hk} = M_{hk} Y_{hk}$ est la production totale du k -ième segment de la strate h , et $C\{\bar{Y}_r, \bar{M}_n\}$ est la covariance de \bar{Y}_r et de \bar{M}_n .
En ce qui a trait aux échantillons de taille fixe avec probabilités inégales, l'estimateur $\bar{Y}_r (= \bar{y})$ est à peu près conditionnellement non biaisé pour le rendement moyen des $n = \sum n_h$ segments de l'échantillon de la première phase. Le rendement moyen des n segments est

$$\bar{\bar{y}}_n = \bar{M}_n^{-1} \sum_{h=1}^H \sum_{n_h} W_h n_h^{-1} Y_{hk}.$$

Par conséquent, la covariance de \bar{y}_r et \bar{M}_n est la covariance de $\bar{M}_n^{-1} \bar{Y}_n$ et \bar{M}_n , où

$$\bar{Y}_n = \sum_{h=1}^H \sum_{n_h} W_h n_h^{-1} Y_{hk}.$$

À l'aide de la formule d'approximation couramment utilisée pour les rapports, la covariance de \bar{y}_r et \bar{M}_n est donnée, approximativement, par

$$C\{\bar{M}_n^{-1} \bar{Y}_n, \bar{M}_n\} \doteq C\{(\bar{Y}_n - \bar{y}_n \bar{M}_n) \bar{M}_n^{-1}, \bar{M}_n\}$$

$$= \bar{M}_n^{-1} [C\{\bar{Y}_n, \bar{M}_n\} - \bar{y}_n V\{\bar{M}_n\}]. \tag{3.19}$$

Si la probabilité que l'on observe le couple (Y_{hk}, M_{hk}) est proportionnelle à π_{hk}^* , l'estimateur de la covariance de \bar{Y}_n et de \bar{M}_n est défini par

$$C\{\bar{Y}_n, \bar{M}_n\} = \sum_L^L W_h^2 n_h^{-1} S_{MYh}, \tag{3.20}$$

où

$$S_{MYh} = K_h (K_h^{-1})^{-1} \left(\sum_{k=1}^{K_h} \pi_{hk}^{*-1} \right)^{-1} \sum_{k=1}^{K_h} \pi_{hk}^{*-1} (M_{hk} - \bar{M}_h) (M_{hk} \bar{Y}_{hk} - \bar{y}_h^*),$$
$$\bar{M}_h^* = \left(\sum_{k=1}^{K_h} \pi_{hk}^{*-1} \right)^{-1} \sum_{k=1}^{K_h} \pi_{hk}^{*-1} M_{hk},$$
$$\bar{y}_h^* = \left(\sum_{k=1}^{K_h} \pi_{hk}^{*-1} \right)^{-1} \sum_{k=1}^{K_h} \pi_{hk}^{*-1} M_{hk} \bar{Y}_{hk}.$$

L'estimateur S_{MYh} équivaut à un estimateur par quotient d'Horvitz-Thompson de la moyenne des produits $(M_{hk} - \bar{M}_h)(Y_{hk} - \bar{Y}_{hk})$ rajusté pour tenir compte du nombre de degrés de liberté. Le facteur de rajustement $K_h(K_h - 1)^{-1}$ est rendu nécessaire parce que la construction du produit exige que l'on remplace les moyennes de la population par les moyennes d'échantillon.
En substituant les équations (3.15), (3.16) et (3.20) aux termes correspondants dans l'équation (3.18), on obtient

$$V\{\bar{Y}_r\} = N^2 [M_n^2 V\{\bar{Y}_r\} + 2\bar{Y}_r C\{\bar{Y}_n, \bar{M}_n\} - \bar{y}_r^2 V\{\bar{M}_n\}], \tag{3.21}$$

où

$$\bar{y}_{hk} = m^{-1} \sum_{\ell=1}^{\ell} y_{hk\ell} \quad \text{si } A_{hk} > 0, \\ \bar{y}_{hk} = 0 \quad \text{si } A_{hk} = 0,$$

$$\bar{M}_r = \sum_L \sum_{h=1}^{h=1} \sum_{k=1}^{k=1} \pi_{hk}^{*-1} M_{hk}.$$

Dans l'équation (3.15) et les autres équations données dans le reste de cette section, \bar{M}_r (nombre total d'unités secondaires) équivaut à A_r (superficie totale); le lecteur peut à son gré substituer l'un à l'autre.

Dans l'analyse qui suit, nous supposons que l'échantillonnage de segments avec remise selon une probabilité proportionnelle à la superficie ensemencée de la culture en question est une approximation de l'échantillonnage systématique avec probabilité proportionnelle à la taille effectué à la seconde phase. Suivant l'hypothèse de l'échantillonnage avec remise, un estimateur de \bar{y} est défini comme étant

$$\bar{V}(\bar{y}_r) = \bar{M}_r^{-2} \sum_L \sum_{h=1}^{h=1} K_h (K_h - 1)^{-1} \sum_{k=1}^{k=1} (\pi_{hk}^{*-1} u_{hk} - \bar{u}_h)^2, \tag{3.16}$$

où

$$u_{hk} = M_{hk} (\bar{y}_{hk} - \bar{y}_r),$$

$$\bar{u}_h = K_h^{-1} \sum_{k=1}^{k=1} \pi_{hk}^{*-1} u_{hk}.$$

Un estimateur de la production totale est défini par

$$\bar{y}_r = N \bar{M}_r \bar{y}_r, \tag{3.17}$$

où

$$\bar{M}_n = \sum_L \sum_{h=1}^{h=1} W_h m_h^{-1} \sum_{k=1}^{k=1} M_{hk}.$$

N est le nombre total de segments dans la population et $W_h = N^{-1} N_h$. La formule d'approximation de Taylor appliquée à la variance non conditionnelle de la distribution approximative de \bar{y}_r est

$$V\{\bar{y}_r\} = N^2 [\bar{M}_N^2 V\{\bar{y}_r\} + 2 \bar{M}_N \bar{y}_N C\{\bar{y}_r, \bar{M}_n\} + \bar{y}_N^2 V\{\bar{M}_n\}], \tag{3.18}$$

où \bar{y}_r est défini en (3.15), \bar{M}_n est défini en (3.17),

$$\bar{M}_N = N^{-1} \sum_L \sum_{h=1}^{h=1} M_{hk},$$

$$\bar{y}_N = \left(\sum_L \sum_{h=1}^{h=1} M_{hk} \right)^{-1} \sum_L \sum_{h=1}^{h=1} \sum_{k=1}^{k=1} Y_{hk},$$

proportionnelle à sa probabilité conditionnelle de sélection à la seconde phase, étant donné les segments échantillonnés à la première phase.

Soit π_{hk} la probabilité conditionnelle que le segment k de la strate h soit prélevé dans la seconde phase, étant donné l'échantillon de segments de la première phase. Nous avons

(3.12)
$$\pi_{hk} = \min(1, M_{hk} \pi_{hkt}) ,$$

où N_{hkt} est une constante dans le segment hk . Si $\pi_{hk} = 1$ et que le segment est échantillonné pour recevoir plus d'une unité secondaire, on suppose que les unités secondaires sont prélevées séparément.

Soit π_{hk}^* la probabilité non conditionnelle que le segment k de la strate h fasse l'objet d'une observation dans la seconde phase. Si $A_{hk} = 0$ alors π_{hk}^* est la probabilité non conditionnelle que le segment hk soit échantillonné pour recevoir au moins une unité secondaire. Si $A_{hk} = 0$ alors π_{hk}^* est égale à la probabilité que le segment hk soit échantillonné dans la première phase d'échantillonnage. Posons

(3.13)
$$\pi_{hk}^* = \frac{N_h}{n} \quad \text{si } A_{hk} = 0 ,$$
$$\pi_{hk}^* = \pi_{hk} \frac{N_h}{n} \quad \text{si } 0 < \pi_{hk} < 1 ,$$

où π_{hk} , définie en (3.12), est la probabilité conditionnelle que le hk -ième segment soit échantillonné à la seconde phase, étant donné l'échantillon de segments de la première phase. Dans notre analyse, nous supposons que π_{hk}^* est fixe. Cette hypothèse se vérifiera et π_{hk}^* sera la vraie probabilité non conditionnelle si π_{hk} est un multiple déterminé de M_{hk} , qui aura été déterminé avant l'échantillonnage. L'expression (3.13) sera une approximation si π_{hk} est une fonction des segments prélevés au premier degré d'échantillonnage.

L'expression (3.13) est proportionnelle à M_{hk} pour $M_{hk} \pi_{hkt} \leq 1$. Si $M_{hk} \pi_{hkt} > 1$, le nombre d'unités secondaires échantillonnées est égal ou supérieur à 1. $M_{hk} \pi_{hkt}$ est le nombre exact d'unités secondaires qu'il faut attribuer aux segments pour conserver un échantillon auto-pondéré d'unités secondaires. Il n'y a jamais plus qu'une unité d'écart entre $M_{hk} \pi_{hkt}$ et le nombre d'unités secondaires effectivement observées par suite d'un échantillonnage systématique avec probabilité proportionnelle à la taille.

Pour simplifier le reste des calculs, nous supposons que l'échantillonnage systématique ne comporte aucune erreur d'arrondissement. En d'autres termes, nous supposons que le nombre d'unités secondaires observées par segment est exactement le nombre nécessaire pour obtenir un échantillon auto-pondéré. Par conséquent, nous supposons que le nombre d'unités secondaires observées dans un segment prélevé à la seconde phase d'échantillonnage est

(3.14)
$$m_{hk} = 1 \quad \text{si } 0 < \pi_{hk} < 1 ,$$
$$m_{hk} = M_{hk} \pi_{hkt} \quad \text{si } \pi_{hk} = 1 .$$

Suivant cette hypothèse, un estimateur par quotient combiné du rendement moyen avec probabilités égales équivaut à l'estimateur défini en (3.1). L'estimateur par quotient combiné est

(3.15)
$$y_r = \hat{M}_r^{-1} \sum_L^h \sum_{K_h}^k \pi_{hk}^{*-1} M_{hk} y_{hk} ,$$

À cause de la simplicité de l'expression (3.8), on a proposé de l'utiliser comme estimateur de la variance non conditionnelle. On a également proposé d'estimer la variance de l'estimation de la production totale d'un Etat par l'équation suivante:

(3.9)
$$V^*(Y) = A^2V_2(Y) + Y^2V(A) + V(A)V_2(Y),$$

où A est défini en (3.6) et V(A) est l'estimateur habituel de la variance d'un total estimé stratifié,

(3.10)
$$V(A) = \sum_L^h N_h^2 n_h^{-1} (n_h - 1)^{-1} \sum_{n_h}^k (A_{hk} - \bar{A}_h)^2,$$

et

$$\bar{A}_h = n_h^{-1} \sum_{n_h}^k A_{hk}.$$

L'estimateur (3.9) est un estimateur de la variance d'un produit, qui est fondé sur l'hypothèse implicite que Y et A sont non corrélés.
Il est difficile d'évaluer dans quelle mesure l'estimateur (3.9) tend à sous-estimer la variance de Y. La variance non conditionnelle de Y est

$$V(Y) = V_1 \{ E_2(Y) \} + E_1 \{ V_2(Y) \}$$

(3.11)
$$= V_1 \{ A^{-1} \sum_L^h N_h n_h^{-1} \sum_{n_h}^k Y_{hk} \} + E_1 \{ V_2(Y) \},$$

où $Y_{hk} = M_{hk} Y_{hk}$, est le total pour le segment k dans la strate h, et E_1 et V_1 désignent respectivement l'espérance mathématique et la variance suivant le plan d'échantillonnage de la première phase.

Dans le cas d'un échantillonnage aléatoire simple des unités secondaires, l'estimateur $V_2(Y)$ est non biaisé pour le second terme du membre de droite de l'équation (3.11). Comme le plan de sondage du NASS prévoit un échantillonnage systématique dans la seconde phase, $V_2(Y)$ est un estimateur biaisé de $V_2(Y)$. La nature et l'importance du biais dépendent de la structure de corrélation des unités de la liste servant à l'échantillonnage de la seconde phase. L'estimateur de la variance $V_2(Y)$ est également biaisé parce qu'on a établi la formule de la variance en supposant un échantillonnage avec remise dans la seconde phase. Dans la mesure où l'échantillonnage à la seconde phase se fait sans remise (comme les échantillons sont prélevés systématiquement dans la liste des segments élargis, seuls les segments de grande superficie sont échantillonnés plus d'une fois), $V_2(Y)$ surestimerait $V_2(Y)$.
L'estimateur $V^*(Y)$ ne renferme pas d'estimateur de $A^2 V_1 \{ E_2(Y) \}$; cela a pour effet de créer un biais négatif. Ce terme n'est toutefois pas facile à estimer, même suivant l'hypothèse simplificatrice d'un échantillonnage avec probabilité proportionnelle à la taille dans la seconde phase. Aussi, nous analyserons l'efficacité de $V^*(Y)$ par la méthode de Monte Carlo (section 3.4).

3.3 Autres estimateurs de variance

Une autre façon d'estimer $V(Y)$ est de considérer l'échantillon comme le résultat d'un sondage à deux phases (voir tableau 1) et de supposer que la probabilité non conditionnelle qu'un segment soit prélevé à l'intérieur d'une strate pour recevoir une unité secondaire est

pour les unités secondaires incluses dans les segments dont $A_{hk} > 0$ où N_h est le nombre global de segments inclus dans la strate h , M_{hk} est le nombre de segments inclus dans la strate h qui ont été prélevés à la première phase. La probabilité conditionnelle que l'on observe un segment non commencé de la culture en question dans la seconde phase est un.

On peut alors exprimer l'estimateur de la moyenne défini en (3.1) par l'équation suivante:

$$\bar{y} = \frac{\sum_{h=1}^L \sum_{k=1}^{K_h} \sum_{\ell=1}^{n_h} \pi_{hk\ell}^{-1} Y_{hk\ell}}{\sum_{h=1}^L \sum_{k=1}^{K_h} \sum_{\ell=1}^{n_h} \pi_{hk\ell}^{-1} \delta_{hk\ell}}, \quad (3.5)$$

où $N_h n_h^{-1}$ est l'inverse de la probabilité de sélection à la première phase, K_h est le nombre de segments prélevés dans la strate h à la seconde phase, et $K = \sum K_h$. Étant donné une échelle appropriée, le numérateur de (3.5) est un estimateur de la production totale et le dénominateur est un estimateur de la superficie totale. Il est possible de montrer que le numérateur est un estimateur sans biais en calculant les espérances mathématiques; pour cela, on définit tout d'abord une relation conditionnelle basée sur les unités échantillonnées de la première phase, puis on fait la moyenne pour l'ensemble de ces échantillons. Le dénominateur est un estimateur stratifié du nombre total d'unités secondaires. Par la nature du sondage, le nombre d'unités d'échantillonnage est proportionnel à la superficie, et on peut choisir l'échelle de telle sorte que le nombre d'unités secondaires soit égal à la superficie. Ainsi, on peut considérer \bar{y} comme le rapport d'un estimateur sans biais de la production totale d'une culture à un estimateur sans biais de la superficie totale consacrée à cette culture.

Pour estimer la production totale d'un État, le NASS multiplie \bar{y} par A , où A est l'estimateur de la superficie cultivée totale, qui prend la forme

$$A = \sum_L \sum_{h=1}^{n_h} N_h n_h^{-1} \sum_{k=1}^{K_h} A_{hk}. \quad (3.6)$$

L'estimation de la production totale est donc

$$Y = A \bar{y}. \quad (3.7)$$

3.2 Estimateurs de variance simples

Suivant l'hypothèse d'un échantillonnage aléatoire simple d'unités secondaires parmi toutes les unités secondaires disponibles dans la seconde phase, l'estimation de la variance conditionnelle de y , étant donné les segments de la seconde phase, est

$$V_2(y) = D^{-1} (D - 1)^{-1} \sum_D^t (Y_t - \bar{y})^2, \quad (3.8)$$

où l'indice inférieur 2 (par rapport à V) sert à identifier la variance conditionnelle et l'indice inférieur t (par rapport à Y) remplace l'indice triple hkl . La somme pour t variant de 1 à D est la somme effectuée pour les D unités secondaires incluses dans les segments ensemble- cés de la culture en question.

Dans la section 3.1, nous étudions l'estimateur de rendement actuellement en usage et voyons à quelles conditions cet estimateur est sans biais pour le rendement moyen d'un Etat. Dans la section 3.2, nous analysons un estimateur simple de la variance du rendement estimé tandis que, dans la section suivante, nous définissons des estimateurs des variances non conditionnelles des estimateurs du rendement et de la production. Enfin, la section 3.4 décrit une simulation de Monte Carlo appliquée aux estimateurs.

3.1 Estimateur de rendement et estimateur de production actuellement en usage

À l'heure actuelle, on estime le rendement moyen d'un Etat comme si l'échantillon était un échantillon aléatoire simple d'unités secondaires avec probabilités égales. L'estimateur est le rendement moyen simple des unités secondaires qui sont ensemencées de la culture considérée. En d'autres termes, l'estimation du rendement moyen à l'acre est définie comme étant

$$(3.1) \quad \bar{y} = D^{-1} \sum_L \sum_{h=1}^h \sum_{k=1}^k \sum_{\ell=1}^{\ell} X_{hke} \delta_{hke},$$

où

$$\delta_{hke} = 1 \quad \text{si } A_{hk} > 0, \\ \delta_{hke} = 0 \quad \text{si } A_{hk} = 0,$$

$$(3.2) \quad D = \sum_L \sum_{h=1}^h \sum_{k=1}^k \sum_{\ell=1}^{\ell} \delta_{hke},$$

m_{hk} est le nombre d'unités secondaires prélevées dans le segment hk , L est le nombre de strates et X_{hke} est l'estimation du rendement à l'acre pour l'unité secondaire ℓ du segment hk . Si la superficie cultivée dans un segment (A_{hk}) est nulle, alors $m_{hk} = 1$ et $X_{hke} = 0$, par définition. Le nombre total d'unités secondaires observées dans les segments ensemencés de la culture en question est D .

L'équation (3.1) peut être réécrite sous la forme opérationnelle suivante:

$$(3.3) \quad \bar{y} = D^{-1} \sum_{\ell=1}^{\ell} Y_{\ell},$$

où l'indice inférieur ℓ remplace l'indice triple hke et où la sommation porte sur les unités secondaires comprises dans les segments ensemencés de la culture en question.

L'estimateur du rendement moyen à l'acre (3.1) est un genre d'estimateur par quotient combiné. On peut vérifier cela en se servant des probabilités de sélection conditionnelles pour réécrire \bar{y} . Dans le plan de sondage du NASS, les segments sont prélevés systématiquement avec probabilités proportionnelles à la superficie accrue et ceux dont la superficie accrue est suffisamment élevée sont inclus à coup sûr dans l'échantillon. Le nombre d'unités secondaires attribuée à cette catégorie de segments est proportionnel à la taille du segment, mise à part l'erreur d'arrondissement. L'arrondissement est effectué suivant le plan d'échantillon-nage systématique. Soit π_{hke} la probabilité conditionnelle que l'unité secondaire ℓ du segment hk de la strate h soit échantillonnée, étant donné l'échantillon de segments prélevés à la première phase d'échantillonnage. Nous avons

3. MÉTHODES D'ESTIMATION

En principe, l'échantillon de l'enquête objective sur le rendement est le résultat d'un sondage à deux phases avec sous-échantillonnage dans la seconde phase. Le tableau 1 donne une description schématique de cet échantillon. L'échantillon produit à la première phase est un échantillon aléatoire simple stratifié de segments. L'échantillon produit à la seconde phase comprend tous les segments qui ne sont pas ensemencés de la culture étudiée de même qu'un échantillon de segments ensemencés de cette culture, ces segments étant prélevés selon une probabilité proportionnelle à la superficie cultivée. L'échantillon de segments est tiré des champs ensemencés de la culture étudiée qui ont été prélevés dans la première phase utilisant l'échantillonnage systématique avec probabilités de sélection proportionnelles à la superficie. Un échantillon d'unités secondaires – où chaque unité secondaire correspond à une paire de parcelles – est ensuite prélevé dans les segments ensemencés qui ont été échantillonnés à la seconde phase.

Puisque l'unité secondaire est toujours une paire de parcelles, nous ne parlerons plus désormais de parcelles mais bien d'unités secondaires. De même, nous oublierons que les champs sont les unités opérationnelles qui servent à délimiter les parcelles et nous parlerons uniquement de segments échantillonnés. Notons que l'on observe deux genres de segments dans la seconde phase; les segments qui sont ensemencés de la culture étudiée et ceux qui ne le sont pas. Le nombre total de segments dans la seconde phase est K . La superficie et la production totale d'un segment qui n'est pas ensemencé de la culture considérée sont connues (toutes deux nulles). En ce qui a trait aux segments ensemencés de la culture considérée, on utilise un sous-échantillon d'unités secondaires pour estimer la production.

Soit M_{hk} le nombre d'unités secondaires incluses dans le segment k de la h -ième strate. On peut supposer, sans perte de généralité, que M_{hk} est égal à A_{hk} (A_{hk} étant la superficie cultivée du segment hk). La validité de cette identité repose essentiellement sur le choix d'une échelle d'équivalence appropriée pour la superficie.

Tableau 1

Méthode d'échantillonnage pour l'enquête objective sur le rendement

Phase/unité d'échantillonnage	Méthode de sélection	Nombre ¹ d'échantillonnées	Données recueillies
Unité primaire d'échantillonnage: segment	probabilité égale à l'intérieur des strates	n_h	superficie cultivée
Unité primaire d'échantillonnage: segment	probabilité inégale	K_h	superficie cultivée, production ² estimée
Unité secondaire d'échantillonnage: paires de parcelles	probabilité égale	m_{hk}	production des parcelles estimée

¹ Par strate pour les unités primaires d'échantillonnage et par segment pour les unités secondaires d'échantillonnage.
² La production du segment est nulle si la superficie ensemencée est nulle, autrement elle est estimée à l'aide des données relatives aux parcelles.

laquelle se trouve l'unité de base. En Californie, par exemple, la taille de segment recherchée est ½ mille carré pour les strates de vergers et 1 mille carré pour toutes les autres strates de terres labourables. Une unité de base contient normalement de 1 à 30 segments aréolaires. Chaque strate d'exploitation est sous-stratifiée en fonction du lieu géographique. Pour former les sous-strates géographiques, on classe les unités de base de chaque strate d'exploitation par comité de manière que des comités voisins qui ont des caractéristiques agricoles comparables sont fondus en un seul (Fecso 1978), puis on forme de nouveaux groupes de segments aréolaires en suivant l'ordre de classement des unités de base. Ainsi, chaque sous-strate renferme des segments aréolaires qui ont les mêmes caractéristiques agricoles et qui sont voisins. Dans une strate d'exploitation donnée, les sous-strates comptent toutes le même nombre de segments et ont toutes la même superficie, après arrondissement. Pour obtenir des renseignements détaillés sur la conception de la base aréolaire, le lecteur est invité à consulter Fecso et Johnson (1981) et Houseman (1975).

Pour l'estimation de la variance, les sous-strates servent de strates d'échantillonnage. Aussi, les sous-strates d'exploitation seront désignées les strates. La première étape de l'échantillonnage consiste à prélever des unités de base dans chaque strate. Le nombre d'unités de base prélevées dépend de la nature de la strate. En règle générale, on prélève de 8 à 15 unités dans les strates de terres labourables tandis qu'on en prélève 4 ou 5 dans les strates agro-urbaines, urbaines et non agricoles. Les unités sont prélevées aléatoirement selon une probabilité proportionnelle au nombre de segments aréolaires qu'elles contiennent. La seconde étape consiste à prélever au hasard un segment aréolaire dans chaque unité de base échantillonnée. Ainsi, la probabilité de sélection est la même pour tous les segments aréolaires d'une même strate.

Dans le plan de sondage qui nous intéresse, l'unité de base est l'unité primaire d'échantillonnage. Comme les unités de base sont prélevées selon une probabilité proportionnelle au nombre de segments qu'elles contiennent et qu'un segment est prélevé dans chaque unité de base échantillonnée, on peut considérer le segment comme l'unité primaire d'échantillonnage. Dans notre étude, les deux premiers degrés de l'échantillonnage ne font qu'un et l'échantillon de segments est considéré comme un échantillon aléatoire simple stratifié à un seul degré. Comme la fraction de sondage moyenne est d'environ 1%, nous n'utiliserons pas le terme correctif de la population finie dans notre analyse.

Les troisième et quatrième degrés de l'échantillonnage consistent respectivement à échantillonner des champs et à prélever des parcelles à l'intérieur de ces champs. Dans l'enquête énumérative de juin, on examine chaque segment aréolaire échantillonné pour y recenser les champs qui ont été ensimencés de la plante d'intérêt ou qui doivent l'être. Les champs ainsi recensés sont classés par numéro de segment et par ordre d'énumération à l'intérieur de chaque segment. On prélève ensuite un échantillon systématique de champs selon une probabilité proportionnelle au produit de la superficie du champ par l'inverse de la probabilité de sélection du segment aréolaire qui contient le champ. Ainsi, le nombre de champs échantillonnés par segment varie et il se peut que les grands champs soient échantillonnés plus d'une fois.

Au quatrième et dernier degré de l'échantillonnage, on définit deux parcelles de superficie comparables dans chaque champ échantillonné à l'aide d'une méthode de délimitation aléatoire fondée sur les rangées et les pas. Lorsque les rangées sont difficiles à distinguer l'une de l'autre ou lorsqu'il s'agit de blé, on délimite les parcelles à l'aide d'un nombre aléatoire de pas en bordure du champ et d'un nombre aléatoire de pas vers l'intérieur du champ. Une méthode différente s'applique aussi dans les enquêtes objectives sur le rendement du blé. Pour ce genre d'enquête, on délimite aléatoirement la première parcelle puis on situe la seconde par rapport à la première. Si un grand champ est échantillonné plus d'une fois au troisième degré de l'échantillonnage, on prélève séparément des paires de parcelles additionnelles. Comme les parcelles sont toujours échantillonnées par paire, une paire de parcelles est désignée ici l'unité secondaire. On ne peut avoir plus de huit parcelles (c'est-à-dire quatre unités secondaires) par champ.

Propriétés statistiques des estimateurs de la production végétale

CAROL A. FRANCISCO, WAYNE A. FULLER, et RON FEECO¹

RÉSUMÉ

Le National Agricultural Statistics Service du Département de l'agriculture des États-Unis fait des enquêtes de rendement pour diverses grandes cultures aux États-Unis. Bien que les méthodes d'échantillonnage des champs varient selon les cultures, le plan de sondage demeure le même pour toutes. Cet article présente une analyse de ce plan de sondage et des estimateurs actuellement utilisés. Les auteurs définissent également d'autres estimateurs du rendement et de la production ainsi que des estimateurs de la variance des estimateurs, puis les comparent aux estimateurs courants par une approche théorique et une simulation de Monte Carlo.

MOTS CLÉS: Enquêtes sur les cultures; estimation du rendement; échantillon à deux phases; estimation de variance.

1. INTRODUCTION

Le National Agricultural Statistics Service (anciennement le Statistical Reporting Service) du Département de l'agriculture des États-Unis procède à des enquêtes objectives sur le rendement du maïs, du coton, du soja, du sorgho à grains, du tournesol et du blé dans les principaux États producteurs. Certains autres pays réalisent le même genre d'enquêtes. Bien que les méthodes d'échantillonnage des champs varient selon les cultures (taille des parcelles, méthodes de délimitation des parcelles et techniques de mesure appliquées aux légumes et aux fruits), le plan de sondage est le même pour toutes les enquêtes. En effet, les enquêtes objectives sur le rendement reposent sur une méthode d'échantillonnage à quatre degrés. On trouvera à la section 2 une description détaillée de ce plan de sondage. Dans la section 3, nous décrivons et évaluons les estimateurs du rendement moyen des cultures et les estimateurs de la variance. Nous y examinons également d'autres estimateurs. La section 4 renferme les conclusions de l'étude et des recommandations.

2. PLAN DE SONDAGE DE L'ENQUÊTE OBJECTIVE SUR LE RENDEMENT

Les deux premiers degrés de l'échantillonnage du plan de sondage produisent l'échantillon de segments aréolaires qui sert à l'enquête énumérative de juin du National Agricultural Statistics Service (NASS). Dans chaque État, la base aréolaire (base de sondage) est stratifiée suivant le mode d'exploitation du sol. L'État de Californie, par exemple, est divisé en 12 strates d'exploitation. Chacune de ces strates est subdivisée en territoires appelés unités de base. La superficie des unités de base varie; la taille réelle d'une unité de base dépend des renseignements qui existent sur la désignation des frontières, des renseignements complémentaires recueillis, des frontières politiques, etc.. Une fois que les unités de base sont définies, on détermine le nombre de segments aréolaires dans chacune, en divisant la superficie totale d'une unité par la taille de segment désirée. Celle-ci varie selon la strate d'exploitation dans

¹ Carol A. Francisco, Syntex Laboratories Inc., 3401 Hillview Avenue, Palo Alto, Californie 94304; Wayne A. Fuller, Département de statistique, Iowa State University, Ames, Iowa 50011; et Ron Feecho, Survey Research Branch, National Agricultural Statistics Service, U.S. Department of Agriculture, Washington, D.C. 20250.

STATISTIQUE CANADA (1987). *Tableaux sommaires pour RTA urbaines et codes postaux ruraux*. No. de catalogue 17-602, Statistique Canada.

TROTTIER, I., et CHOUDHRY, G.H. (1985). Model based unemployment estimates for small areas. Dans *Small Area Statistics, An International Symposium ('85 Contributed Papers)*, (éds. R. Platek et M.P. Singh), Laboratory for Research in Statistics and Probability, Carleton University/Université d'Ottawa.

collaboration étroite avec les organismes fournisseurs est extrêmement important. La capacité de l'organisme statistique d'influer sur la conception ou la refonte de systèmes administratifs repose sur une compréhension mutuelle des besoins des deux organismes intéressés. Si l'on pouvait adopter, à l'échelle de l'appareil gouvernemental, une politique ou un principe dominant à Statistique Canada, la façon de concevoir de la conception des systèmes administratifs ou, plus généralement, de la façon de satisfaire les besoins statistiques des nouveaux programmes, cela faciliterait la tâche de l'organisme statistique mais ne remplacerait pas la création d'un climat de collaboration étroite avec les organismes administratifs.

Il est possible d'envisager divers mécanismes qui permettraient à l'organisme statistique d'avoir accès aux dossiers administratifs et d'influer sur leur conception à l'intérieur de l'appareil gouvernemental. L'applicabilité et l'efficacité de chaque mécanisme dépendront du contexte législatif et politique ainsi que du mandat et du statut de l'organisme statistique au sein de l'appareil gouvernemental. D'après l'expérience de Statistique Canada, l'établissement de relations de travail bilatérales étroites avec les ministères administratifs, fondées sur le principe des avantages réciproques, est l'approche la plus efficace. Il importe que l'utilisation des dossiers administratifs reçoive un appui politique, et cet appui s'est manifesté dans les décisions récentes du gouvernement relativement aux compressions budgétaires.

REMERCIEMENTS

Pour cette communication, nous nous sommes inspirés des idées et des travaux d'un bon nombre de personnes à Statistique Canada. Nous y avons notamment incorporé des passages extraits du document rédigé en collaboration avec Michael Colledge, Ivan Fellegi et John Leyes.

BIBLIOGRAPHIE

AUGER, E. (1987). Family data from the Canadian personal income tax file. Article présenté à la réunion annuelle de 1987 de l'American Statistical Association.

BRACKSTONE, G.J. (1984). Répercussions des changements technologiques sur le recensement. *Estadística*, 36, 43-60.

CANADA, Loi sur la statistique (1971). *Statuts du Canada 1970-71-72*, c.15.

CANADA, Loi sur la protection des renseignements personnels (1982). *Statuts du Canada 1980-81-82*, c.11.

CHOUDHRY, G.H., et HIDIROGLOU, M.A. (1987). Small area estimation: Some experiences at Statistics Canada. *Proceedings of the 46th Session of the International Statistical Institute*, (en voie de rédaction).

COLLEDGE, M.J. (1987). The Business Survey Redesign Project: Implementation of the new strategy at Statistics Canada. *Proceedings of the Third Annual Research Conference, US Bureau of the Census*, (en voie de rédaction).

FEENEY, G.A. (1987). The estimation of the number of unemployed at the small area level. Dans M.P. Singh, New York: John Wiley.

NORRIS, D.A., et STANDISH, L.D. (1983). A technical report on the development of migration data from tax records. Document de travail, Statistique Canada.

ROWEBOTTOM, L.E. (1978). The utilization of administrative records for statistical purposes. *Survey Methodology*, 4, 1-15.

f) Il faudrait que les moyens pris pour assurer la protection matérielle dont l'utilisation des dossiers administratifs délicats fait l'objet soient évidents, et cette protection devrait peut-être même être plus grande que celle qui est habituellement exercée au Bureau. À Statistique Canada, les divisions qui sont les principales détentrices des données fiscales sont logées dans des locaux à accès limité, situés dans des bâtiments à l'entrée desquelles on effectue des vérifications de sécurité.

g) Le fait que les fichiers statistiques ne sont pas soumis à l'examen des services de sécurité ou de renseignement permet en grande partie d'entretenir le sentiment de confiance que le public éprouve à l'égard de la confidentialité absolue des données fournies à l'organisme statistique. Lorsque le Service canadien du renseignement de sécurité a été créé en 1983, Statistique Canada a obtenu une telle exemption pour ses données (la seule accordée à un organisme fédéral).

Les points que nous venons de mentionner correspondent à des mesures précises qui peuvent être prises pour éviter que le public ne réagisse mal à l'utilisation des dossiers administratifs ou pour contrer une telle réaction une fois qu'elle a eu lieu, mais ce dont l'organisme statistique a surtout besoin, c'est d'un appui politique fort. Le capital politique à gagner d'une réduction visible des coûts et du fardeau de réponse, conjugué à des garanties politiques fermes de protection des données des particuliers, constitue une base solide sur laquelle les politiciens peuvent s'appuyer pour dissiper les inquiétudes du public au sujet de l'utilisation des dossiers administratifs à des fins statistiques. En même temps, ils doivent démentir immédiatement et de façon non équivoque toute suggestion selon laquelle les dossiers statistiques sont utilisés à des fins administratives.

8. CONCLUSION

Les dossiers administratifs sont et demeureront une source de plus en plus utile de données statistiques. Les points forts et les points faibles relatifs des données tirées des systèmes administratifs, du point de vue des coûts, de l'univers visé, et de leur qualité, utilité et actualité par rapport aux données du recensement ou aux données d'enquête dictent la façon la plus efficace d'utiliser ces différentes sources de données. Entre autres utilisations courantes, il y a les totalisations directes, les estimations indirectes, le remplacement des réponses d'enquête, la constitution et la mise à jour de bases de sondage et l'évaluation de données. Ces utilisations sont aujourd'hui assez répandues dans la plupart des programmes statistiques et elles le seront probablement encore davantage à l'avenir.

Au Canada, les dossiers administratifs font à présent partie intégrante de l'appareil statistique. C'est en partie grâce à leur utilisation que Statistique Canada a pu maintenir ses programmes malgré la compression de ses budgets. Du fait même, le fardeau de réponse a été réduit et on a commencé à produire de nouvelles séries de données ou à produire les anciennes plus souvent. Comme nous ne disposons pas de registres administratifs, il nous a fallu porter une attention très particulière aux questions de couverture et d'utilisation combinée de données administratives et de données d'enquête pour nous assurer que les estimations de totaux pour un univers soient valides. Le recours à des techniques de jumelage d'enregistrements, bien que nécessitant un contrôle rigoureux, s'est avéré d'une très grande utilité, notamment en ce qui a trait aux données sur les entreprises, aux études longitudinales du marché du travail et aux enquêtes épidémiologiques.

En raison de leur utilisation croissante des dossiers administratifs, les organismes statistiques sont de plus en plus tributaires des autres organismes pour la fourniture continue des données dont ils ont besoin pour leurs programmes. Quels que soient les textes de loi et les politiques qui régissent les activités de l'organisme statistique, l'établissement d'ententes de

assurer par la suite), les études longitudinales pour lesquelles on utilise des dossiers administratifs, les enquêtes épidémiologiques et enfin les études d'évaluation visant à vérifier les réponses dans des questionnaires d'enquête en les comparant à des données de sources administratives. Le fait de devoir expliquer pourquoi on a besoin d'identificateurs alors que l'identité des personnes auxquelles les données se rapportent ne présente aucun intérêt est un défi de taille pour l'organisme statistique.

Une autre source de préoccupation a trait à l'engagement même de respecter la confidentialité. Même si Statistique Canada a toujours été irréprochable en matière de protection de la confidentialité, il y a sans doute des répondants qui sont sceptiques quant à la protection accordée aux renseignements qui les concernent. Il est possible que l'utilisation de recenseurs connus des répondants, en particulier dans les petites localités, ajoute à leur scepticisme. Certains répondants semblent occuper comme principe qu'il se fait beaucoup d'échanges de renseignements entre les ministères fédéraux et, dans certains cas, ne font de distinction entre les différents ministères.

Une autre inquiétude encore aurait trait non pas à la confiance dont seraient dignes les services ayant actuellement la garde de banques de renseignements, mais découlerait de la crainte de n'avoir aucune protection contre une violation éventuelle de la confidentialité, soit par des moyens illégaux soit par des personnes élues qui auraient un point de vue différent sur la protection des renseignements personnels. Pour protéger le public contre cette éventualité, il faudrait qu'on élimine des bases de données statistiques toute information permettant d'identifier les personnes dont il y est question.

Cette préoccupation du public au sujet de la protection et du traitement des renseignements personnels oblige l'organisme statistique à penser à des mesures susceptibles de prévenir ou de réduire les réactions défavorables du public à l'utilisation légitime des dossiers administratifs pour des travaux statistiques. Comme il s'agit là essentiellement d'une question de perception du public, il importe que l'organisme statistique ne cache rien de ses pratiques, et que le public soit bien au courant des mesures en application, notamment :

- a) L'information publique destinée aux répondants et aux utilisateurs doit toujours mettre l'accent sur l'importance accordée à la confidentialité de toutes les données individuelles (microdonnées) acquises par l'organisme statistique.
- b) Le fait que les microdonnées ne circulent que dans un sens devrait être souligné. L'organisme statistique reçoit des microdonnées, mais il n'en ressort que des données agrégées ou résumées, dont la confidentialité est garantie par leur nature. Cela est vrai aussi bien pour les données de recensement ou les données d'enquête que pour les données tirées de dossiers administratifs.

- c) Il faudrait faire valoir les avantages qui découlent de l'utilisation des dossiers administratifs : réduction du fardeau de réponse et allègement du fardeau fiscal pour les contribuables. Ces affirmations devraient s'appuyer sur des chiffres et des faits.
- d) Des règles explicites en matière de jumelage d'enregistrements énonçant les conditions dans lesquelles l'organisme statistique peut exercer cette activité permettrait, d'une part, de montrer que le jumelage d'enregistrements ne se fait ni à la légère ni sans contrôle et, d'autre part, de ne pas donner suite aux demandes de jumelage d'enregistrements qui ne rempliraient pas les conditions requises.

- e) En vertu de la Loi sur la protection des renseignements personnels, les individus doivent être informés des fins auxquelles les renseignements personnels les concernant sont recueillis. Il faudrait inciter les organismes administratifs à indiquer de façon précise l'utilisation statistique dont ces renseignements pourraient faire l'objet. Même si l'utilisation de données à des fins statistiques constitue une utilisation secondaire acceptable des dossiers administratifs, la mention explicite de ces fins sur le formulaire de collecte permettrait d'éviter les mauvaises surprises.

7. CONFIDENTIALITE, PROTECTION DES RENSEIGNEMENTS PERSONNELS ET RELATIONS AVEC LE PUBLIC

Même si l'organisme statistique a le pouvoir légal d'exploiter les dossiers administratifs et même si l'organisme administratif fournit ces dossiers dans un esprit de grande collaboration, il faut faire très attention à la perception que le public a de l'utilisation des dossiers administratifs à des fins autres que celles pour lesquelles ces dossiers ont été conçus. Comme l'efficacité, sinon la survie, d'un organisme statistique dépend en très grande partie du maintien de la collaboration et de la confiance des répondants, l'organisme en question ne doit pas s'engager à la légère dans une activité qui pourrait saper cette collaboration ou cette confiance. Dans beaucoup de pays, le public est, depuis quelques années, de plus en plus sensible à tout ce qui concerne la protection des renseignements personnels et, par conséquent, à l'accès à l'information et au contrôle de cet accès. Au Canada, l'adoption de la Loi sur la protection des renseignements personnels en 1982 témoigne de cette préoccupation grandissante. Cette loi dit, entre autres choses et avec certaines exceptions, qu'un répertoire de toutes les banques de renseignements personnels relevant d'institutions gouvernementales fédérales doit être publié périodiquement, que toutes les personnes ont le droit d'avoir accès aux renseignements les concernant contenus dans ces banques de renseignements et que les renseignements personnels ne peuvent servir qu'à des usages qui sont compatibles avec les fins pour lesquelles ils ont été obtenus. Une des exceptions prévues à cette dernière disposition est que des renseignements personnels peuvent être divulgués

... à toute personne ou à tout organisme, pour des travaux de recherche ou de statistique, pourvu que... les fins auxquelles les renseignements sont communiqués ne peuvent être normalement atteintes que si les renseignements sont donnés sous une forme qui permette d'identifier l'individu qu'ils concernent, (et) la personne ou l'organisme (s'engageant) par écrit (après du responsable de l'institution) à s'abstenir de toute communication ultérieure des renseignements tant que leur forme risque vraisemblablement de permettre l'identification de l'individu qu'ils concernent." (Loi sur la protection des renseignements personnels 1982 alinéa 8(2) (j)).

Cette disposition régit l'utilisation des dossiers administratifs à des fins statistiques du point de vue de la Loi sur la protection des renseignements personnels. L'article en question est toutefois formulé sous réserve de toute autre loi du Parlement, de sorte que si une disposition d'une loi régissant le processus administratif interdisait une telle utilisation, celle-ci aurait préséance. Bien qu'il soit reconnu dans la Loi sur la protection des renseignements personnels et dans d'autres lois que les travaux statistiques constituent, dans certaines conditions, une utilisation secondaire légitime des dossiers administratifs, cela ne dissipe pas pour autant les préoccupations du public au sujet d'une utilisation possible des banques de données qui irait à l'encontre des intérêts des particuliers. Il est douteux que les citoyens moyens fassent la différence entre l'utilisation statistique, où l'identité du titulaire de chaque dossier ne présente plus aucun intérêt, et l'utilisation administrative, où il est essentiel de savoir à qui chaque dossier se rapporte. Il serait plus facile d'expliquer cette différence et de s'en servir si l'on pouvait dire sans équivoque que, pour le genre de travaux qui sont faits en statistique, on n'a jamais besoin d'identificateurs. Malheureusement, ce n'est pas le cas. Plusieurs techniques statistiques tout aussi légitimes exigent qu'on se serve d'identificateurs au cours des manipulations intermédiaires des données. Ces techniques supposent toutes une forme quelconque d'appariement des données provenant de différents fichiers ou de différentes sources, et les identificateurs permettent d'apparier les bons dossiers ensemble. Une fois cette opération effectuée, les dossiers peuvent être rendus anonymes si aucun couplage subséquent n'est prévu. Citons, à titre d'exemples de cas où une forme d'identification est nécessaire, le recensement de la population (pour veiller à ce que la couverture soit exhaustive et s'en

- (iii) offrir des conseils ou des services techniques au service statistique de l'organisme administratif;
- (iv) adopter une politique officielle en matière de collecte de renseignements exigeant que tout projet de collecte de données (dans un but statistique ou administratif) soit soumis à l'examen d'un organisme central;
- (v) demander que chaque nouvelle proposition de programme soit assortie d'un plan prévoyant la façon d'obtenir les renseignements statistiques nécessaires pour contrôler et évaluer le programme en question;
- (vi) promouvoir l'utilisation des définitions statistiques standard (p. ex., ce que l'on entend par les termes "famille", "établissement commercial" ou "chômeur") dans les systèmes administratifs;
- (vii) faire en sorte que les vérificateurs des activités gouvernementales recommandent l'utilisation des dossiers administratifs comme moyen économique par excellence de collecte de renseignements;
- (viii) suivre une orientation politique favorisant un recours accru à certains systèmes administratifs ou la recherche de solutions de échange aux enquêtes;
- (ix) supprimer les obstacles législatifs qui limitent l'accès aux dossiers administratifs ou à leur utilisation à des fins statistiques.

L'expérience vécue par le Bureau dans ses relations avec les autres ministères fédéraux a été particulièrement fructueuse lorsque des ententes bilatérales étroites ont été conclues. La création de comités supérieurs bilatéraux au début des années 80 a favorisé et parfois assuré l'élaboration de telles ententes. Les mesures prises à l'échelle de l'appareil gouvernemental, par exemple la gestion de l'information et la planification statistique, ont moins bien réussi à favoriser l'utilisation des dossiers administratifs. Les vérifications des opérations gouvernementales et les directives du cabinet ont bien donné une certaine impulsion à des activités visant à accroître l'utilisation des données administratives, mais l'augmentation de cette utilisation dépend elle-même, encore une fois, de l'existence de liens de travail étroits avec certains ministères. Bien qu'il convienne de décrire l'organisme statistique comme un organisme progressif qui tente de briser les barrières irrationnelles faisant obstacle à l'utilisation des données administratives, il faut quand même également reconnaître qu'il puisse exister au sein de l'organisme même une certaine résistance au changement. Les membres du personnel dont les carrières ont été vouées à la conception et à la réalisation d'enquêtes peuvent avoir besoin d'arguments convaincants pour admettre que les restrictions budgétaires et les besoins en données nous obligent à présenter à combiner les enquêtes et d'autres méthodes. Comme les remarques qui ont précédé s'appliquent uniquement aux systèmes administratifs par des ministères fédéraux, il convient d'ajouter quelques mots au sujet des dossiers provinciaux. Certaines des mesures mentionnées conviendraient aussi aux dossiers administratifs des provinces; toutefois les rapports avec les administrations infranationales posent un problème fondamental, celui de la conformité à des normes communes. La différence entre les besoins et les priorités des provinces, accentuée par des possibilités technologiques croissantes, entraînera une diversité de plus en plus grande des systèmes administratifs s'il n'existe pas de force centralisatrice. Dans le passé, le Bureau a eu recours à divers mécanismes pour essayer de favoriser une certaine uniformisation, mais avec plus ou moins de succès. Comme dans le cas des services fédéraux qui ont la garde de dossiers administratifs, l'avantage réciproque doit être l'argument principal. Il y a des comités fédéraux-provinciaux dans plusieurs domaines. Le Conseil de la statistique de l'état civil, composé d'officiers provinciaux de l'état civil et de représentants du Bureau, est un exemple d'une longue et fructueuse collaboration. Des comités comme celui-là ont, dans le passé, élaboré des conventions pour la déclaration de certains éléments de données et ils en ont suivi l'application. Par exemple, le système de déclaration des finances municipales a été conçu à la suite de réunions fédérales-provinciales sur les statistiques financières des municipalités.

6.2 Droit de regard sur les changements

Nous avons déjà mentionné l'incidence que des modifications apportées aux pratiques ou règlements administratifs auraient sur les statistiques produites. Il suffirait qu'on change la clientèle d'un programme, qu'on introduise une mesure incitant les personnes visées par un programme à y participer ou au contraire à ne plus y participer ou bien qu'on modifie la procédure de manière à altérer la qualité ou l'exhaustivité des dossiers, et les séries chronologiques seraient interrompues. L'organisme statistique doit donc être sur ses gardes et réagir chaque fois qu'un changement impose de l'extérieur risque de se produire.

Il y a, en revanche, des changements que l'organisme statistique aimerait bien voir se réaliser. Le statisticien qui souhaite utiliser des dossiers administratifs éprouve souvent un sentiment de frustration en pensant que ceux-ci seraient beaucoup plus utiles si seulement de petites modifications y étaient apportées. Par exemple, l'ajout d'une nouvelle question, l'utilisation d'une notation différente, l'inclusion de nouveaux sous-groupes ou l'exécution d'une nouvelle vérification de la qualité pourraient améliorer sensiblement la valeur statistique des dossiers. Par contre, pourquoi l'organisme administratif envisagerait-il d'apporter des changements qui ne sont pas requis par le processus administratif, quand ces changements contribueraient probablement dans une certaine mesure à augmenter les coûts et la complexité de ce processus?

L'organisme statistique a donc un défi à relever : celui de persuader les administrateurs que les avantages découlant de ces changements seraient plus grands que les coûts supplémentaires générés. La difficulté vient de ce que ce n'est pas nécessairement le ministère responsable du système administratif qui profite des avantages en question, mais plutôt les ministères décisionnaires ou d'autres utilisateurs qui se servent des données à des fins statistiques.

Il est habituellement plus facile de tenir compte des exigences statistiques au moment de concevoir un système que de modifier un système déjà en cours d'exploitation. Par conséquent, un mécanisme qui permettrait de tenir compte des besoins en données statistiques au stade de la conception ou de la restructuration en profondeur d'un système administratif serait préférable à un mécanisme qui tenterait seulement d'adapter les systèmes en place. En ce moment, au Canada, on parle beaucoup de la réforme de la taxe que le gouvernement songe à instituer. Cette réforme pourrait changer significativement la collecte des données sur les entreprises au Canada. La participation de statisticiens à la conception d'un système comme celui-là ajouterait beaucoup aux avantages statistiques qu'on pourrait en retirer. Bien entendu, l'établissement d'un nouveau système administratif est un événement relativement rare, de sorte qu'il faut également adapter les systèmes en place si l'on veut obtenir des avantages statistiques à court terme. Par ailleurs, le fait qu'il soit relativement exceptionnel qu'on convoque ou qu'on restructure des systèmes administratifs importants renforce l'argument selon lequel il ne faut rater aucune occasion de prendre part à ce genre de travaux.

6.3 Mécanismes

Il existe divers moyens ou mécanismes permettant à l'organisme statistique d'avoir un certain droit de regard sur les systèmes administratifs et d'y avoir accès; certains de ces mécanismes sont essentiellement bilatéraux, mettant en jeu l'organisme statistique et un service administratif quelconque, tandis que d'autres concernent l'ensemble du secteur public. En voici quelques exemples :

- (i) créer des comités bilatéraux constitués de hauts fonctionnaires qui examineraient les questions intéressant les deux organismes, notamment les problèmes relatifs à la fourniture de données administratives;
- (ii) fournir à l'organisme administratif les données statistiques afin de montrer à la fois l'utilité des données et, le cas échéant, les lacunes attribuables aux pratiques administratives;

(éducation, santé, justice) pour la création de bases de sondage et pour l'obtention de données. Les responsables des recherches en cours sur les enquêtes téléphoniques et les registres d'adresses utilisent les dossiers administratifs pour dresser les listes de logements ou de ménages. Une enquête interne récente a dénombré plus d'une cinquantaine de systèmes administratifs actuellement exploités à des fins statistiques. Tous les types de dossiers et toutes les catégories d'utilisation mentionnés aux sections 2 et 3 sont représentés, y compris des registres de maladies et des dossiers renfermant des données sur les véhicules automobiles immatriculés, les mouvements d'aéronefs, la commercialisation du lait, la taxe de vente sur le carburant, les permis de construire municipaux et les droits de douane et d'accise.

6. ACCÈS AUX SYSTÈMES ADMINISTRATIFS ET DROIT DE REGARD SUR LEUR CONTENU

Le tour d'horizon que nous venons de faire de l'utilisation des dossiers administratifs à des fins statistiques montre bien que ces dossiers constituent une source d'information indispensable pour de nombreux programmes de Statistique Canada. Cela nous amène aux mesures que le Bureau peut prendre pour protéger ces sources de données et peut-être rendre ces dernières plus utiles du point de vue statistique. Cette section porte sur les deux principaux problèmes, c'est-à-dire l'accès aux dossiers administratifs et la manière dont on peut influencer leur contenu, leur conception et la marche à suivre à leur égard.

6.1 Accès

Le fondement juridique de l'accès aux dossiers administratifs est l'article 12 de la Loi sur la statistique (1971) qui se lit comme suit:

Une personne ayant la garde ou la charge de documents ou archives conservés dans un département ou dans un bureau municipal, une corporation, entreprise ou organisation et dont on pourrait tirer des renseignements que l'on cherche à obtenir pour les objets de la présente loi ou qui aideraient à compléter ou à corriger ces renseignements, doit en permettre l'accès, à ces fins, à une personne autorisée par le statisticien en chef à obtenir ces renseignements ou cette aide pour le complètement ou la correction de ces renseignements.

Cette disposition, qui semble donner un droit assez étendu à l'accès, comporte cependant des restrictions. Dans certains cas, les lois régissant le processus administratif limitent l'accès aux données administratives ou leur utilisation secondaire. Il en résulte une incompatibilité entre les lois qui, au mieux, a pour effet de retarder les négociations au sujet de l'accès. Dans d'autres cas, l'accès à des fins statistiques est permis de façon explicite.

L'adoption de lois facilitant l'accès aux dossiers administratifs est une condition nécessaire mais non suffisante de l'utilisation productive des dossiers administratifs. Il serait probablement beaucoup plus efficace, lorsqu'on cherche à obtenir l'accès aux dossiers administratifs, d'essayer d'y arriver par la voie de la collaboration (sur le plan de l'élaboration et de l'utilisation de ces dossiers dans un but statistique) que d'avoir recours à des dispositions ou à des sanctions légales. En effet, une fois l'accès obtenu, il est possible de parvenir à l'étape suivante (qui consiste à influencer sur la conception ou les méthodes d'utilisation) seulement s'il existe un esprit de collaboration entre l'organisme administratif et l'organisme statistique.

L'accès aux dossiers administratifs dont jouit le Bureau est strictement un phénomène à sens unique. Les microdonnées individuelles vont de l'organisme administratif à l'organisme statistique, et seulement des données agréées, protégées par la confidentialité, peuvent retourner en sens inverse. La seule exception technique à cette règle se produit lorsque l'organisme administratif dépend de l'organisme statistique pour organiser, préparer, mettre en forme, traiter ou restructurer ses dossiers et que les microdonnées originales sont renvoyées sous une autre forme à l'organisme fournisseur.

Les fichiers des déclarations d'impôt des particuliers sont également utilisés pour produire les estimations de la migration annuelle. Pour cela, les dossiers d'une même personne se rapportant à deux années consécutives sont apparés et le code attribué à la division de recensement (ou au comté) dans chaque dossier est comparé. S'il n'est pas le même, on en déduit que le déclarant a déménagé. Les données démographiques et les renseignements relatifs aux exemptions permettent d'estimer le nombre total de personnes qui ont déménagé avec le déclarant. Dans une dernière étape, comme le fichier fiscal ne couvre pas l'ensemble de la population, on procède à un redressement pour produire l'estimation annuelle du nombre total de migrants. Depuis 1981, les chiffres produits en se servant du fichier fiscal sont utilisés dans le cadre du programme des estimations démographiques de Statistique Canada. Norris et Standish (1983) décrivent en détail la méthode utilisée pour estimer la migration au moyen des dossiers fiscaux.

Nous avons vu plus tôt qu'il était possible d'établir des statistiques sur le revenu des particuliers grâce aux données fiscales; le revenu des familles présente encore plus d'intérêt tant du point de vue des possibilités d'analyse que du point de vue des programmes sociaux qui y sont liés. Pour déduire le revenu des familles au moyen des dossiers fiscaux individuels, il faut pouvoir déterminer quels titulaires de dossiers appartiennent à une même famille et ensuite apparier les données qui les concernent. Des travaux de cette nature sont en cours et les résultats sont prometteurs. Pour une description des méthodes employées et des résultats obtenus, voir Auger (1987).

Le système d'assurance-chômage (A.-C.) constitue la deuxième source importante de données administratives sur les particuliers. Statistique Canada a accès au fichier des demandeurs et à celui des bénéficiaires. Ces deux fichiers contiennent les enregistrements de personnes qui, pour diverses raisons, ont peut-être droit aux prestations. Il ne s'agit pas toujours de chômeurs au sens de la définition internationale de chômage retenue par l'enquête sur la population active (EPA), d'où sont tirées les données publiées sur le sujet. S'il y avait moyen de trouver dans le système de l'A.-C. une catégorie se rapprochant de celle de l'EPA, il serait possible de produire à partir de ses fichiers des données sur le nombre de chômeurs dans les petites régions. Cependant, comme dans le meilleur des cas les catégories de l'A.-C. ne sont pas identiques à celles de l'EPA, il faut donc plutôt chercher à harmoniser ou concilier ces deux sources de données. Notamment, les chiffres mensuels de l'A.-C. sur les petites régions pourraient être pris comme indicateurs de la variation du chômage à l'échelon local et ils pourraient ensuite être corrigés en fonction d'estimations fiables de l'EPA correspondant à un niveau géographique supérieur (p. ex. une province). Plusieurs méthodes d'estimation de ce genre ont été examinées (comme la méthode de l'estimation par régression et celle de l'estimation par le quotient avec maintien de la structure), mais il n'a pas encore été possible de déterminer laquelle est la meilleure. Troitier et Choudhry (1985) donnent une description des travaux effectués dans ce domaine et Feeney (1987) parle d'une approche semblable utilisée dans les travaux australiens. L'utilisation de modèles à séries chronologiques exploitant la structure corrélée de l'erreur dans le temps est une méthode qui semble très prometteuse (Choudhry et Hidiroglou 1987).

Les exemples que nous venons de donner montrent qu'on utilise principalement les données tirées des dossiers administratifs sur les particuliers pour effectuer des totalisations directes ou des estimations, contrairement aux données tirées des dossiers administratifs concernant les entreprises, dont on se sert surtout pour mettre à jour la base de sondage ou carrément remplacer les données d'enquête.

Bien que les domaines d'activité auxquels ces deux exemples ont trait soient en pleine expansion et revêtent beaucoup d'importance à Statistique Canada, ils ne touchent qu'une petite fraction du nombre de fichiers administratifs dont on se sert au Bureau. À titre d'exemple, les dossiers administratifs sont depuis longtemps universellement utilisés en statistique sociale

de trouver des solutions de rechange moins onéreuses. Il est vite devenu évident que la meilleure façon de réaliser, au Canada, le potentiel statistique des dossiers administratifs sur les particuliers serait de s'en servir non pas pour remplacer le recensement quinquennal, mais pour le compléter en allant chercher des données pour les petites régions dans ces dossiers pendant la période intercensitaire. Comparativement au recensement, les systèmes de dossiers administratifs actuels ne permettent pas d'obtenir le même niveau de couverture, une aussi grande précision géographique ni un éventail comparable de renseignements sur les caractéristiques des particuliers, des familles et des ménages. On continue néanmoins de chercher à atteindre le même taux de couverture que le recensement en combinant plusieurs systèmes de dossiers administratifs et de voir s'il est possible de remplacer les réponses fournies à certaines questions du recensement par des données tirées de sources administratives.

Nous allons examiner l'utilisation des dossiers administratifs pouvant compléter les données de recensement pendant la période intercensitaire. Les travaux expérimentaux réalisés jusqu'à présent ont porté surtout sur les systèmes administrés à l'échelle nationale (p. ex., l'impôt sur le revenu, l'assurance-chômage (A.-C.), les allocations familiales et la sécurité de la vieillesse) plutôt que sur les systèmes administrés au niveau provincial ou infraprovincial (p. ex., l'assurance-maladie, les permis de conduire et les évaluations municipales). Ces derniers posent, en plus des problèmes propres à l'utilisation de tous les dossiers administratifs, celui du manque d'uniformité entre les secteurs de compétence.

Le fichier des déclarations d'impôt annuelles des particuliers (T1) est la principale source de statistiques sur les Canadiens. Au départ, on s'en est servi pour des totalisations directes qui ont permis de produire des statistiques sur le revenu et l'activité selon l'âge et le sexe, aux niveaux provincial et infraprovincial. Le lieu de résidence des déclarants est déterminé au moyen du code postal indiqué dans le dossier. Un fichier de conversion permettant d'établir la correspondance entre les codes postaux et les diverses divisions géographiques de recensement (province, comté, municipalité, circonscription électorale, etc.) a été créé. Il est également possible de produire des totalisations spéciales pour les régions définies par les utilisateurs en fonction des codes postaux.

Les données ainsi obtenues sont, bien sûr, fondées sur des notions, des définitions et des règlements découlant de la Loi de l'impôt sur le revenu. Les définitions ne sont donc pas nécessairement les mêmes que celles qui permettent d'effectuer les analyses requises (p. ex., il n'y a pas de données sur certains régimes d'aide sociale qui ne sont pas assujettis à l'impôt). Le revenu peut être ventilé selon la source, par exemple le revenu d'emploi peut être séparé. Les variables pouvant faire l'objet d'un classement recoupé sont limitées; (p. ex., l'âge, le sexe et l'état matrimonial). Bien qu'il y ait une question sur la profession dans le formulaire de déclaration d'impôt, la qualité des réponses et de leur codage n'est pas assez bonne pour qu'on puisse s'en servir à des fins statistiques. La couverture se limite aux personnes qui sont tenues de produire une déclaration d'impôt : les personnes à faible revenu et les personnes à charge sont donc sous-représentées. Il se peut que les modifications apportées à la législation fiscale aient, avec le temps, une incidence marquée sur le taux de couverture, lequel a nettement augmenté, entre 1977 et 1978, lorsque il a été décidé que les personnes à faible revenu devraient faire une déclaration si elles voulaient obtenir le nouveau crédit d'impôt pour enfants.

Malgré toutes ces réserves, la totalisation directe des données tirées des fichiers fiscaux permet de produire, pendant la période intercensitaire, des données pour les petites régions sur le revenu qui sont fort utiles. Une publication récente de Statistique Canada fait état de statistiques ainsi établies pour les régions de tri d'acheminement, c.-à-d., les régions regroupant les adresses dont les trois premiers caractères du code postal sont les mêmes (Statistique Canada 1987). Étant donné qu'une principale préoccupation, lorsqu'on publie des données pour les petites régions, est de s'assurer qu'il est impossible d'en déduire des renseignements concernant des particuliers, on ne donne pas de statistiques pour les régions où moins de 100 personnes ont fait des déclarations d'impôt.

de retenue sur la paye (RP) de Revenu Canada est une source importante de renseignements sur les changements qui surviennent dans les entreprises (naissances et réorganisations). La mise en activité d'un nouveau compte de RP par un employeur est le signal qu'il y a du nouveau. On procède alors à un suivi qui permet de savoir si une mise à jour s'impose. Les autres signaux proviennent des déclarations d'impôt annuelles, des réponses données aux enquêtes normales et de l'établissement routine des profils.

On n'essaie pas de déterminer les unités qui composent les petites entreprises ni le lien qui existe entre ces unités, car il y a trop de mouvement dans ce secteur. On se sert carrément des données administratives. Deux listes d'entreprises peuvent servir de base de sondage pour les enquêtes: la première est établie à partir des déclarations d'impôt annuelles les plus récentes et la seconde, à partir des comptes actifs de RP. Toutes les entreprises dont la taille dépasse le seuil fixé sont enlevées de ces listes. Ces dernières se superposent et, pour chaque enquête, on se sert de celle qui convient le mieux. Pour les enquêtes infra-annuelles, on privilégie la liste établie à partir des comptes de RP, qui est plus à jour car il est possible d'ouvrir ou de fermer un compte à n'importe quel moment de l'année. Par contre, on n'y trouve pas les entreprises qui n'ont pas d'employés.

4.2 Remplacement des données d'enquête

Afin de réduire le fardeau de réponse et les coûts, on remplace autant que possible les données d'enquête par les données fiscales. Les notions de base et les définitions sur lesquelles ces dernières sont fondées ne coïncident pas toujours avec celles que les enquêtes doivent utiliser pour des raisons d'uniformité avec le système de comptabilité nationale ou pour les besoins d'analyses. Il faut donc prendre soin d'extraire des déclarations d'impôt seulement les éléments d'information qui correspondent le plus aux définitions de l'enquête. En outre, les données fiscales ne couvrent pas l'ensemble de variables que de nombreuses enquêtes-entreprises cherchent à étudier. Notamment, on n'y trouve pas de statistiques sur la production.

Un autre problème que pose l'utilisation des données fiscales concerne l'établissement du lien qui existe entre l'unité pour laquelle on a une déclaration d'impôt et l'unité ou les unités visées par l'enquête. On rencontre ce problème particulièrement dans le cas des grosses entreprises à structure complexe dont il a été question plus tôt.

La stratégie élaborée relativement aux enquêtes annuelles consiste à utiliser les données fiscales principalement pour les petites entreprises où il y a, en général, une déclaration d'impôt par unité d'exploitation. Le fardeau de réponse des petites entreprises est ainsi nettement réduit, et la qualité des données finales n'en souffre pas trop puisqu'on obtient la plus grande partie des renseignements dont on a besoin sur l'activité économique au moyen des déclarations que les grandes entreprises font dans le cadre des enquêtes.

Il ressort clairement de ce bref aperçu des nouvelles stratégies et infrastructure établies pour les enquêtes-entreprises que la poursuite de ce programme est fondamentalement tributaire des données fiscales. Statistique Canada et Revenu Canada doivent donc travailler en collaboration très étroite afin que l'incidence de toute réforme, administrative ou autre, du système fiscal puisse être évaluée avant qu'elle n'entre en vigueur et afin que les mesures nécessaires soient prises à temps.

5. DONNÉES SOCIO-ÉCONOMIQUES TIRÉES DE DOSSIERS ADMINISTRATIFS

À la fin des années 70, on a commencé à faire un effort concerté pour utiliser les données sur les particuliers, les familles et les ménages qui se trouvaient dans les dossiers administratifs. Cette démarche était motivée à l'origine par la croissance du coût des enquêtes et la nécessité

4. DONNÉES ADMINISTRATIVES ET ENQUÊTES-ENTREPRISES

Un projet de remaniement complet de l'infrastructure et de la stratégie sur lesquelles s'appuie le programme d'enquêtes-entreprises est actuellement en cours à Statistique Canada. Il s'agit en particulier de procéder à la refonte du registre d'entreprises (qui sert de base de sondage aux enquêtes-entreprises), de repenser le rôle et l'utilisation des données fiscales et d'élaborer une stratégie cohérente pour la conception d'enquêtes-entreprises tenues une ou plusieurs fois par an. Ce remaniement a été entrepris car il fallait:

- (a) éliminer les faiblesses évidentes qui existent actuellement sur le plan de la qualité des données;
- (b) harmoniser les données provenant de différentes enquêtes;
- (c) réduire au minimum le fardeau de réponse en utilisant autant que possible les données fiscales;
- (d) réduire la quantité de ressources nécessaires à la mise à jour des bases de sondage des enquêtes.

Colledge (1987) donne une description détaillée de ce projet.

Les données fiscales et celles qui sont extraites des comptes de retenues sur la paye jouent un rôle primordial dans la réalisation des enquêtes-entreprises. En vertu de la Loi sur la statistique, Statistique Canada a accès aux déclarations d'impôt annuelles des sociétés (T2) et des particuliers (T1). Les retenues d'impôts que les employeurs font sur la paye de leurs salariés sont également à sa disposition. Ces données remplissent deux fonctions distinctes à Statistique Canada:

- (i) elles servent à la mise à jour de la base de sondage (liste d'entreprises);
- (ii) elles remplacent les données concernant l'impôt sur le revenu qui pourraient être recueillies au moyen d'une enquête.

4.1 Mise à jour de la base de sondage

La mise à jour de la liste d'entreprises est une tâche complexe pour deux raisons principales: premièrement, la structure d'un grand nombre d'entreprises de même que les liens qui existent entre elles sont loin d'être simples et, deuxièmement, il est difficile de se tenir au courant du nombre élevé de "naissances" et de "morts" qui surviennent dans le secteur des petites entreprises. La définition du terme "entreprise" est elle-même une tâche délicate. Il faut établir une distinction entre la structure juridique (entreprise constituée en société), la structure opérationnelle (comment l'entreprise est organisée pour mener ses activités) et la structure statistique (les unités sur lesquelles on a besoin de données pour effectuer les analyses). À l'intérieur de chaque structure, il est possible d'établir une hiérarchie. Ainsi, la structure statistique se compose, de haut en bas, des entreprises, des sociétés, des établissements et des emplacements. Pour tenir à jour la base de sondage, il ne suffit pas d'y ajouter les nouvelles unités (naissances) et de supprimer celles qui ont mis fin à leurs activités (morts); il faut également tenir compte des changements qui se produisent dans les liens entre les unités des entreprises à structure complexe, et cela veut aussi dire le lien entre la hiérarchie statistique et la hiérarchie opérationnelle.

La stratégie proposée consisterait à faire une mise à jour continue de la structure des entreprises qui dépassent une certaine taille (le seul variant selon la branche d'activité) et du lien entre cette structure et les unités déclarantes du point de vue de l'impôt. L'activité économique des entreprises touchées représenterait environ 70% de l'activité du secteur auquel elles appartiennent. L'"établissement du profil" est un travail qui permet de déterminer la structure interne des entreprises complexes. Il consiste à interviewer des dirigeants d'entreprise pour prendre connaissance de la structure opérationnelle et établir les unités statistiques adéquates. Le système

Comparaison des recensements, des enquêtes et des dossiers administratifs
comme sources de données statistiques

Tableau 1

Éléments		Recensements	Enquêtes	Dossiers administratifs
1. Couverture	Visent à une couverture à 100% de la population	Certaines enquêtes ne comprennent pas certains secteurs de la population (par exemple, les réserves indiennes, les régions éloignées)	Couvrent habituellement un petit nombre de sujets mais permettent de façon plus approfondie qu'un recensement	Établis en fonction des exigences administratives
2. Contenu	La grande diversité des éléments de données permet beaucoup de classements recoupés	Peuvent être établis en fonction des besoins de l'analyse sociale et économique	Peuvent être produits dans la plupart des cas	Peuvent être produits, pourvu que chaque dossier ait un code géographique correspondant à une petite région
3. Concepts/définitions	Disponibles par suite des efforts déployés pour arriver à une couverture complète	Étant de plus petite taille que les recensements, ils permettent un contrôle encore plus serré que dans les recensements	Relativement faible par enquête, même si le coût cumulé d'une enquête habituelle portant sur une période intermédiaire de cinq ans peut être élevé	Peuvent être annuels ou mensuels selon le programme administratif
4. Estimations relatives aux petites régions	Peut être conçu pour réduire au minimum les erreurs	Coût relativement faible par enquête, même si le coût cumulé d'une enquête habituelle portant sur une période intermédiaire de cinq ans peut être élevé	Peuvent être annuels, trimestriels ou mensuels selon les sujets	Peuvent être annuels ou mensuels selon le programme administratif
5. Contrôle qualitatif	Très élevé	Les données peuvent être obtenues six mois à 2 1/2 ans après le jour du recensement	Les enquêtes habituelles répétées produisent des résultats en quelques semaines, alors qu'il se peut qu'une forme utilisable d'un fichier annuel ne soit pas prête avant qu'une bonne partie de l'année suivante soit écoulée	Des changements peuvent être apportés en raison de modifications législatives ou de modifications de règlements ou en raison de modifications des pratiques administratives
6. Coût	Tous les 5 ou 10 ans (selon les sujets)	Les données peuvent être obtenues six mois à 2 1/2 ans après le jour du recensement	Peuvent être annuels, trimestriels ou mensuels selon les sujets	Peuvent être annuels ou mensuels selon le programme administratif
7. Fréquence	Les changements sont sous le contrôle des statisticiens qui répondent aux besoins des utilisateurs	Les changements sont sous le contrôle des statisticiens qui répondent aux besoins des utilisateurs	Dans des enquêtes répétées, les changements sont peu fréquents pour permettre les comparaisons temporelles	Des changements peuvent être apportés en raison de modifications législatives ou de modifications de règlements ou en raison de modifications des pratiques administratives
8. Actualité des données	Les données peuvent être obtenues six mois à 2 1/2 ans après le jour du recensement	Les données peuvent être obtenues six mois à 2 1/2 ans après le jour du recensement	Les enquêtes habituelles répétées produisent des résultats en quelques semaines, alors qu'il se peut qu'une forme utilisable d'un fichier annuel ne soit pas prête avant qu'une bonne partie de l'année suivante soit écoulée	Des changements peuvent être apportés en raison de modifications législatives ou de modifications de règlements ou en raison de modifications des pratiques administratives
9. Stabilité	Les changements sont sous le contrôle des statisticiens qui répondent aux besoins des utilisateurs	Les changements sont sous le contrôle des statisticiens qui répondent aux besoins des utilisateurs	Dans des enquêtes répétées, les changements sont peu fréquents pour permettre les comparaisons temporelles	Des changements peuvent être apportés en raison de modifications législatives ou de modifications de règlements ou en raison de modifications des pratiques administratives
10. Fardeau du répondant	Lourd mais peu fréquent; on le réduit par l'utilisation d'échantillonnage	Lourd mais peu fréquent; on le réduit par l'utilisation d'échantillonnage	Léger en moyenne, bien que lourd pour ceux qui ont été choisis	Aucun fardeau supplémentaire

les petites entreprises au lieu de la collecte des données requises au moyen d'une enquête. Les données fiscales, corrigées s'il y a lieu, sont combinées aux données d'enquête sur les grandes entreprises pour produire des données agrégées sur la branche d'activité.

On classe également dans cette catégorie d'utilisation le couplage de différents fichiers administratifs ou statistiques dans le but de produire des estimations. Par exemple, le couplage du registre des décès avec les fichiers des personnes exposées à certains risques dans le but d'estimer les taux de mortalité différentielle, ou encore le couplage de dossiers fiscaux avec des dossiers de l'assurance-chômage et de la formation de la main-d'oeuvre dans le but d'analyser la participation et l'adaptation au marché du travail.

(3) Bases de sondage

Dans cette catégorie, nous incluons l'utilisation de dossiers administratifs pour créer, compléter ou mettre à jour les bases de sondage servant aux recensements ou aux enquêtes. Un bon exemple serait l'utilisation des renseignements que les employeurs fournissent à Revenu Canada au sujet des retenues sur la paye. Le questionnaire que les titulaires d'un nouveau compte de retenues sur la paye doivent remplir permet de savoir exactement si de nouvelles entreprises ont été créées ou si des modifications ont été apportées à la structure d'entreprises existantes. Même s'il n'y a pas au Canada de registre de logements, un deuxième exemple serait l'utilisation des permis de bâtir ou des listes de nouveaux abonnés au téléphone ou au service d'électricité pour déterminer qu'un logement est nouveau.

(4) Évaluation des enquêtes

Cette catégorie comprend l'utilisation de dossiers administratifs à des fins de vérification, de validation ou d'évaluation de données produites au moyen d'une enquête, et ce au niveau de chaque unité ou à un niveau agrégé. Dans le passé, on s'est servi des dossiers de l'immigration et de l'impôt pour évaluer les questions du recensement sur l'immigration et les revenus, et des dossiers des allocations familiales pour vérifier le taux de couverture du recensement en ce qui a trait aux enfants.

Il y a un facteur important qui permet de déterminer l'utilisation qui sera faite d'une source de renseignements administratifs, et c'est la qualité présumée des données administratives comparativement à celle des données d'enquête correspondantes. Certaines études d'évaluation utilisent les dossiers administratifs pour évaluer les réponses aux questions des enquêtes, alors que d'autres se servent de données d'enquête comme données de référence pour évaluer les estimations produites à partir de données administratives. La qualité des dossiers administratifs doit être évaluée cas à cas. Du point de vue de l'utilisation statistique qu'on va en faire, elle dépend en général d'au moins trois facteurs:

- (i) Les définitions utilisées pour les besoins du système administratif;
- (ii) Le taux de couverture voulu du système administratif;
- (iii) La qualité des données recueillies et de leur traitement par les services administratifs.

Si un de ces trois facteurs présente des faiblesses, l'utilité statistique des dossiers administratifs risque d'être compromise. La rapidité avec laquelle ces dossiers peuvent être obtenus est un autre aspect important à considérer. Les contraintes qui peuvent éventuellement limiter cette utilisation et qu'il faut par conséquent considérer avant toute décision ont été traitées ailleurs (voir par exemple Brackstone, 1984). Le tableau 1 compare les points forts et les points faibles des dossiers administratifs et ceux des recensements et des enquêtes. Pour illustrer l'utilisation qui est faite au Canada des dossiers administratifs, nous allons décrire deux domaines d'application à Statistique Canada. Le premier a trait à la production de statistiques sur les entreprises et le second, à la production de statistiques sur les particuliers et les familles.

(6) *Les dossiers liés à la prestation de services publics.*

Ces services comprennent l'électricité, le téléphone et l'eau. Le taux de couverture des abonnés et la qualité des renseignements relatifs à la prestation des services et à la facturation sont bons en général. Plusieurs de ces services sont administrés au niveau provincial ou municipal.

La façon dont les dossiers administratifs sont constitués varie également beaucoup d'un organisme à un autre. La plupart des systèmes administratifs dont l'univers visé est grand sont à présent automatisés, mais le fait que le matériel utilisé et la présentation des données diffèrent non seulement d'un secteur de compétence à un autre mais aussi entre l'organisme administratif et l'organisme statistique pose un problème qu'il reste à régler. L'automatisation de plus en plus répandue entraîne également le problème des nombreuses modifications apportées aux dossiers initiaux par l'organisme administratif intéressé avant qu'ils soient transmis à l'organisme statistique. Alors qu'un meilleur contrôle de la qualité des formulaires reçus aurait peut-être de bonnes chances d'améliorer la qualité finale du fichier administratif, l'organisme statistique est obligé, lui, de procéder à des travaux supplémentaires pour connaître et évaluer l'incidence de toute opération de traitement préliminaire que l'organisme administratif a pu exécuter. Dans certains services administratifs, les dossiers individuels demeurent à l'endroit où ils ont été constitués et seules les données agrégées sont conservées dans un système central. Cette pratique restreint la capacité de l'organisme statistique d'évaluer la qualité des données et limite les possibilités en matière d'analyse.

Enfin, les dossiers diffèrent du point de vue de leur accessibilité. Il y a des lois et des règlements qui régissent l'accès à certains dossiers administratifs ainsi que leur utilisation à des fins secondaires, c'est-à-dire autres que celles pour lesquelles ils ont été constitués, notamment à des fins statistiques. Nous reparlerons de cette question à la section 6.

3. UTILISATION DES DOSSIERS ADMINISTRATIFS

Les données administratives peuvent se prêter à quatre catégories d'utilisation statistique. La plupart des applications qui sont faites de ces dossiers à des fins statistiques se rattachent à une de ces catégories, en représentent une combinaison ou en constituent une variante.

(1) *Totalisations directes*

On inclut ici le comptage des unités dans les fichiers, leur classement recoupé selon certaines caractéristiques et le regroupement des variables quantitatives associées à chaque unité. Les statistiques sur l'état civil et le commerce extérieur sont des exemples importants. Mentionnons également la publication des chiffres mensuels sur le nombre de demandeurs et de bénéficiaires de l'assurance-chômage ventiles selon la province, l'âge, le sexe, la durée des prestations et leur nature ainsi que la production des sommaires annuels sur la répartition du revenu dans chaque comté d'après les données du fichier de l'impôt sur le revenu des particuliers.

(2) *Estimations indirectes*

Cette catégorie comprend les travaux où l'on se sert de données tirées des dossiers administratifs pour effectuer une estimation. Un exemple serait le couplage de deux déclarations d'impôt consécutives d'un particulier dans le but de produire des estimations partielles de la migration, lesquelles peuvent être pondérées à l'aide des données de recensement utilisées comme données de référence. Ces estimations de la migration sont ensuite utilisées dans le programme d'estimations démographiques de Statistique Canada, au même titre que les données administratives sur les naissances, les décès et l'immigration. Un deuxième exemple serait l'utilisation des données fiscales sur

2. TYPES DE DOSSIERS ADMINISTRATIFS

Des dossiers administratifs, il y en a de toutes sortes. Il est important de faire la distinction entre ceux qui sont tenus au niveau national (habituellement par le gouvernement fédéral) et ceux qui le sont à un niveau inférieur (par les provinces ou les municipalités, par exemple). Pour que ces derniers soient utiles au niveau national, il faut que divers secteurs de compétence s'entendent notamment sur les définitions, les normes, la présentation des dossiers et la marche à suivre. Il n'est pas toujours facile de conclure de telles ententes, surtout dans des domaines qui sont constitutionnellement du ressort des provinces.

Les dossiers administratifs diffèrent par leur objet, lequel détermine en grande partie l'univers visé et leur qualité, et par conséquent leur utilité statistique. On peut distinguer six grandes catégories de dossiers administratifs selon leur raison d'être:

- (1) *Les dossiers tenus pour contrôler la circulation des biens et des personnes aux frontières.* Ces dossiers incluent notamment les dossiers sur les importations, les exportations, l'immigration et l'émigration. L'univers visé et le contenu de ces dossiers administratifs dépendent des lois et des règlements à faire respecter et du succès avec lequel ces lois et règlements sont appliqués. Dans l'ensemble, ils sont bien appliqués. Les dossiers administratifs sur l'immigration ne contiennent pas d'information concernant les immigrants illégaux, par définition, mais sont complets à d'autres égards. Toutefois, comme l'émigration n'est pas contrôlée, il n'existe pas de dossiers administratifs concernant l'émigration. Les dossiers administratifs sur les importations canadiennes sont en général plus précis que les dossiers sur les exportations parce que le besoin d'établir les droits de douane frappant les importations requiert l'information plus détaillée.
- (2) *Les dossiers tenus en vertu de l'obligation légale d'enregistrer certains événements.* À titre d'exemple de ces événements, citons les naissances, les décès, les mariages, les divorces, les constitutions d'entreprises en société ou les fusions d'entreprises et l'octroi de licences. De façon générale, l'univers visé et la qualité des dossiers recueillis à cette fin sont très élevés au Canada, car tout citoyen doit fournir une preuve que ce genre d'enregistrement a été fait pour obtenir des droits ou des avantages.
- (3) *Les dossiers nécessaires à la prestation d'avantages et à l'administration d'engagements d'assistance.* Citons par exemple les impôts, l'assurance-chômage, les pensions, l'assurance-maladie et les allocations familiales. L'univers visé et le contenu de ces dossiers dépendent beaucoup des programmes auxquels ils se rapportent. L'univers visé peut être très bien couvert, mais, pour des raisons politiques ou administratives, sa définition risque de ne pas être la plus utile pour les travaux d'analyse.
- (4) *Les dossiers nécessaires à l'administration des établissements publics.* Il s'agit notamment des dossiers concernant les écoles, les universités, les établissements de soins de santé, les tribunaux et les prisons. De façon générale, ces dossiers portent principalement sur le nombre de cas dont l'établissement a la responsabilité et non sur chaque personne qui y fait un séjour. Par contre, ils donnent habituellement des statistiques agrégées très complètes sur la population qui fréquente ces établissements. Au Canada, beaucoup de dossiers administratifs de cette catégorie sont du ressort des provinces.
- (5) *Les dossiers qui découlent de la réglementation de certaines activités économiques par le gouvernement.* Citons les dossiers relatifs au transport, aux affaires bancaires, à la radiodiffusion et aux télécommunications ainsi qu'à la gestion des approvisionnements ou des prix de certaines marchandises, particulièrement dans le secteur agricole.

Utilisation des dossiers administratifs à des fins statistiques

G.J. BRACKSTONE¹

RÉSUMÉ

Les demandes pour les statistiques sur tous les aspects de nos vies, de notre société et de notre économie continuent à augmenter. En même temps, les organismes statistiques, ainsi que beaucoup de leurs répondants, s'inquiètent de plus en plus du fardeau de réponse imposé par les enquêtes. L'utilisation des dossiers administratifs à des fins statistiques s'est présentée comme un des moyens alternatifs pour satisfaire ces demandes statistiques. Cet article expose l'expérience récente à Statistique Canada dans ce domaine, et discute des entraves à une plus grande exploitation de sources administratives. On examine aussi des méthodes possibles pour rendre les systèmes administratifs plus utiles à des fins statistiques, ainsi que certaines questions importantes liées à la protection des renseignements et au jumelage d'enregistrements.

MOTS CLÉS: Estimation indirecte; base de sondage; évaluation des enquêtes; accès; confidentialité.

1. INTRODUCTION

La demande de statistiques se rapportant à de nombreux aspects de notre vie, de notre société, de notre économie et de notre environnement ne cesse d'augmenter. Le fait que nous sommes en mesure, de nos jours, de traiter et de manipuler de vastes ensembles de données, puisque nous entrons dans ce qu'il est convenu d'appeler l'ère de l'information, explique peut-être en partie ce phénomène; celui-ci pourrait également découler de la complexité croissante des systèmes sociaux et économiques et de notre désir de mieux comprendre ces derniers. Quelle que soit la raison qui motive la demande, les organismes statistiques publics doivent y répondre, et ce dans un climat de grandes compressions budgétaires. Ces organismes sont en même temps conscients du fardeau de réponse accru que les répondants auraient à subir s'il fallait mener davantage d'enquêtes pour obtenir les renseignements voulus.

Ces facteurs nous ont amenés à chercher d'autres façons de procéder, la principale étant un plus grand recours aux systèmes administratifs existants. L'idée n'est pas nouvelle. Depuis longtemps déjà les renseignements relatifs à l'état civil, aux importations, aux exportations, aux soins de santé et à l'éducation nous parviennent sous forme de produits dérivés des processus administratifs. Nous verrons plus loin que l'utilisation des données administratives s'est étendue ces dernières années au domaine des statistiques sur les entreprises, les familles et les particuliers.

Les quatre prochaines sections de l'article décrivent les différents types de dossiers administratifs et l'utilisation qui en est faite, notamment à Statistique Canada. Il ressortira de cet exposé que l'appareil statistique canadien est fortement tributaire des dossiers administratifs. La section 6 discute de la question de l'accès aux dossiers administratifs et les moyens de rendre ces dossiers mieux adaptés aux usages statistiques. Enfin, un survol des préoccupations que l'exploitation des dossiers administratifs suscite à l'égard de la protection des renseignements personnels est fourni.

¹ G.J. Brackstone, Statisticien en chef adjoint, Statistique Canada, 26-J Immeuble R.H. Coats, Tunney's Pasture, Ottawa, Ontario K1A 0T6.

- HENSON, R., ROTH, A., et CANNELL, C.F. (1974). Personal vs. telephone interviews and the effects of telephone re-interviews on reporting of psychiatric symptomatology. Document de recherche, Survey Research Center, University of Michigan.
- HOCHSTIM, J.R. (1967). A critical comparison of three strategies of collecting data from households. *Journal of the American Statistical Association*, 62, 976-986.
- HOLT, D., SCOTT, A.J., et EWINGS, P.D. (1980). Chi-squared tests with survey data. *Journal of the Royal Statistical Society, Ser. A*, 143, 303-320.
- IBSEN, C.A., et BALLWEG, J. (1974). Telephone interviews in social research: some methodological considerations. *Quality and Quantity*, 7, 181-192.
- JOWELL, R., et WITHERSPOON, S. (1985). *British Social Attitudes: The 1985 Report*. Aldershot: Gower.
- JORDAN, L.A., MARCUS, A.C., et REEDER, L.G. (1980). Response styles in telephone and household interviewing: a field experiment. *Public Opinion Quarterly*, 44, 210-222.
- KAHN, R.L., et GROVES, R.M. (1977). *Comparing telephone and personal interview systems*. Survey Research Center, University of Michigan.
- KORMENDI, E., EGSMOSE, L., et NOORDHOEK, J. (1986). Datakvalitet ved Telefon-interview. Socialforskningsinstituttet, Studie 52, Copenhagen.
- LOCANDER, W.B., et BURTON, J.P. (1976). The effect of question form on gathering income data by telephone. *Journal of Marketing Research*, 13, 189-192.
- LOCANDER, W.B., SUDMAN, S., et BRADBURN, N. (1974). An investigation of interview method, threat and response distortion. *Proceedings of the Social Statistics Section, American Statistical Association*, 21-27.
- LUCAS, W.A., et ADAMS, W.C. (1977). *An Assessment of Telephone Survey Methods*. Santa Monica, California: Rand Corporation.
- McCULLAGH, P., et NELDER, J.A. (1983). *Generalised Linear Models*. London: Chapman and Hall.
- MILLER, P.V., et CANNELL, C.F. (1982). A study of experimental techniques for telephone interviewing. *Public Opinion Quarterly*, 46, 250-269.
- Market Research Development Fund, (1985). *Comparing telephone and face-to-face surveys*. Marplan Ltd.
- OKSENBERG, L., COLEMAN, L., et CANNELL, C. (1984). Voices and refusal rates in telephone surveys. Document non publié.
- O'NEIL, M., GROVES, R., et CANNELL, C. (1979). Telephone interview intrusions and refusal rates: experiments in increasing respondent cooperation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 252-255.
- ROGERS, T.F. (1976). Interviews by telephone and in person. *Public Opinion Quarterly*, 40, 51-65.
- SCHMIEDESKAMP, J.W. (1962). Reinterviews by telephone. *Journal of Marketing*, 26, 28-34.
- SYKES, W., et HOINVILLE, G. (1985). Telephone interviewing on a survey of social attitudes: a comparison with face-to-face procedures. Social and Community Planning Research Center.
- TODD, J., et BUTCHER, R. (1982). *Electoral Registration in 1981*. London: OPCS.
- WILLIAMS, E. (1977). Experimental comparisons of face-to-face and mediated communication. *Psychological Bulletin*, 84, 963-976.
- WISEMAN, F. (1972). Methodological bias in public opinion surveys. *Public Opinion Quarterly*, 34, 105-108.
- WISEMAN, F., et McDONALD, P. (1979). Noncontact and refusal rates in consumer telephone surveys. *Journal of Marketing Research*, 16, 478-484.

BIBLIOGRAPHIE

- ARONSON, S. (1971). The Sociology of the Telephone. *International Journal of Comparative Sociology*, 12, 153-167.
- BALL, D.W. (1968). Towards a Sociology of Telephones and Telephoners. Dans *Sociology and Everyday Life* (ed. Marcello Truzzi), Englewood Cliffs, New Jersey: Prentice Hall.
- BERGSTEN, J.W. (1979). Some Methodological Results from Four Statewide Telephone Surveys Using Random Digit Dialing. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 239-243.
- BISHOP, Y.M.N., FIENBERG, S.E., et HOLLAND, P.W. (1975). *Discrete Multi-variate Analysis: Theory and Practice*. Cambridge, Mass: MIT Press.
- BLANKENSHIP, A.B. (1977). *Professional Telephone Surveys*. London: McGraw-Hill.
- BLYTH, W.G., et MARCHANT, L.J. (1973). A self-weighting random sampling technique. *Journal of the Market Research Society*, 15, 157-162.
- CANNELL, C.F., OKSENBERG, L., et CONVERSE, J.M. (1979). Experiments in interviewing techniques. Research Report, Institute for Social Research, University of Michigan.
- CHRISTOFFERSEN, M.N. (1984). The quality of data collected at telephone interviews. Danish National Institute of Social Research, Copenhagen.
- COLLINS, M. (1983). Telephone interviewing in consumer surveys. *Market Research Society Newsletter*, octobre.
- COLLINS, M., et SYKES, W. (1987). The problems of non-coverage and unlisted numbers in telephone surveys in Britain. *Journal of the Royal Statistical Society*. Sér. A, 150, (en voie de rédaction).
- COLOMBOTOS, J. (1965). The effects of personal vs. telephone interviews on socially acceptable responses. *Public Opinion Quarterly*, 29, 457-458.
- COLOMBOTOS, J. (1969). Personal versus telephone interviews: effect on responses. *Public Health Reports*, 84, 773-782.
- COOMBS, L., et FREEMAN, R. (1986). Use of telephone interviews in a longitudinal fertility study. *Public Opinion Quarterly*, 28, 112-117.
- CZAJA, R., BLAIR, J., et SEBESTIK, J.P. (1982). Respondent selection in a telephone survey: a comparison of three techniques. *Journal of Marketing Research*, 19, 381-385.
- DE MAIO, T.J. (1984). Refusals in telephone surveys: when do they occur? Document présenté à la 39^{ième} conférence annuelle de l'American Association for Public Opinion Research.
- DILLMAN, D. (1970). *Mail and Telephone Surveys: The Total Design Method*. New York: John Wiley.
- DILLMAN, D., GALLEGOS, J., et FREY, J. (1976). Reducing refusal rates for telephone interviews. *Public Opinion Quarterly*, 40, 66-78.
- FALTHZIK, A. (1972). When to make telephone interviews. *Journal of Marketing Research*, 9, 451-452.
- FITTI, J.E. (1979). Some results from the telephone health interview survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 244-249.
- FLEISHMAN, E., et BERK, M. (1979). Survey of interviewer attitudes towards methodological issues in the national medical care expenditure survey. Document présenté à la troisième conférence bien-nale sur la recherche et la méthodologie relatives aux enquêtes sur la santé, Reston, Virginia.
- GROVES, R., et KAHN, R. (1979). *Surveys by Telephone: A National Comparison with Personal Interviews*. New York: Academic Press.
- GROVES, R.M., MAGILAVY, L.J., et MATHIOWETZ, N.A. (1981). The process of interviewer variability. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 438-443.
- GROVES, R.M., et MATHIOWETZ, N.A. (1984). Computer assisted telephone interviewing: effects on interviewers and respondents. Survey Research Center, University of Michigan.

qui se rapproche le plus de la question sur le revenu dans l'interview sur place: la personne interviewée devait consulter une fiche sur laquelle les tranches de revenu figuraient par ordre croissant. Pour l'interview téléphonique, on énumérait les tranches de revenu en commençant par la plus basse. Les résultats sont indiqués dans le tableau 7.

Dans les différentes études, les enquêtes se sont montrées plus ou moins disposées à répondre à la question sur le revenu suivant le mode d'interview utilisé. Des différences de distribution de réponses, en l'occurrence attribuables à une sous-déclaration probable de revenu, n'ont été observées que dans la première enquête.

3.1.4 Questions complexes

Dans les deux études du SCPR, on a pu préalable déterminer un certain nombre de questions qui pouvaient poser des problèmes de réponse particuliers aux personnes interviewées par téléphone. Ces questions étaient celles qui pouvaient renfermer une ou plusieurs notions difficiles à saisir, les questions longues et les questions à choix multiples. Ces questions "complexes" ne semblent pas poser plus de problèmes aux personnes interviewées par téléphone qu'à celles interviewées sur place. Par exemple, des 19 questions "complexes" définies comme telles dans la première étude (12 de ces 19 questions s'accompagnaient de cartons dans les interviews sur place), une seule a eu des effets attribuables au mode d'interview.

4. SOMMAIRE ET CONCLUSIONS

Comme le nombre de personnes qui ont le téléphone au Royaume-Uni est encore relativement faible, surtout pour certaines couches de la population, il n'est pas probable que l'interview téléphonique puisse bientôt remplacer l'interview sur place pour les enquêtes qui exigent la participation des personnes moins favorisées. Néanmoins, on reconnaît de plus en plus les possibilités qu'offre ce mode d'interview lorsqu'il est combiné à l'interview sur place. Dans l'enquête sur la population active au Royaume-Uni, par exemple, la seconde interview et les interviews subséquentes se font par téléphone dans le cas des répondants admissibles qui ont accepté d'être contactés par téléphone.

Quant il y a utilisation simultanée des deux modes d'interview, il est essentiel pour le succès d'une enquête que cette circonstance ne provoque pas de distorsions. Les résultats que nous venons d'analyser sont encourageants pour l'avenir. À quelques exceptions près, on n'a relevé aucune différence statistiquement significative entre les distributions des réponses obtenues sur place et les distributions des réponses obtenues par téléphone. Néanmoins, les taux de réponse relativement faibles des enquêtes par téléphone posent des problèmes qui doivent être résolus. Des taux de refus élevés peuvent réduire l'efficacité en termes de coûts de l'interview téléphonique. Fait plus important encore, ils augmentent la possibilité d'introduction d'un biais dans l'échantillon. Pour mettre entièrement en valeur la méthode d'interview mixte au Royaume-Uni, il est nécessaire de poursuivre la recherche pour trouver des moyens d'améliorer les taux de réponse des enquêtes par téléphone.

REMERCIEMENTS

Cette étude a été réalisée sous l'égide du Centre des méthodes d'enquête du SCPR, subventionné à titre de centre de recherche désigné par le Conseil de la recherche économique et sociale (subvention HR 3333). Les auteurs tiennent à exprimer leur reconnaissance aux rédacteurs et aux arbitres pour leurs commentaires nombreux et utiles sur les premières versions de cette étude.

- ii) Croyez-vous que dans l'ensemble, la Grande-Bretagne est trop, trop peu ou juste assez gênèreuse envers les Asiatiques et les Antillais établis au pays?
- iii) Enfin, j'aimerais que vous me disiez si vous jugez acceptable qu'une personne homosexuelle occupe un poste de professeur dans une école.

Il n'y avait aucune question qui a donné lieu à des différences significatives dans les distributions de réponse marginales. Pour plusieurs questions, toutefois, les personnes interviewées sur place avaient un peu plus tendance à donner des réponses conformistes au point de vue social. En d'autres mots, les questions semblaient moins délicates au téléphone. Par exemple, 28% des personnes interviewées par téléphone ont déclaré avoir été interrogées par la police relativement à un crime au cours des deux dernières années alors que dans le cas des personnes interviewées sur place, la proportion correspondante a été de 20%.

On a également observé, pour les questions délicates portant sur le crime et la sexualité dans l'enquête du MRDF, que les personnes interviewées par téléphone avaient une faible tendance à donner des réponses plus "honnêtes"; en général, toutefois les différences de distributions observées n'étaient pas significatives. Par exemple, lorsqu'on a demandé aux personnes interviewées de se décrire sur un certain nombre de points, les personnes interviewées par téléphone se sont dites plus "séduisantes" (une moyenne de 2.81 sur 4 comparée à 2.72 sur place) et ils étaient plus enclin à répondre (88% ont répondu comparé à 75% pour les interviews sur place).

Les questions relatives au revenu ont toujours été considérées comme une source de problèmes dans les enquêtes par téléphone (problèmes liés à la volonté de participer des enquêtes et aux réponses fournies). La sous-déclaration du revenu est le problème le plus fréquent, bien qu'en pratique il puisse être difficile de faire la distinction entre la sous-déclaration et la sous-estimation découlant d'un taux de non-réponse plus élevé dans les tranches de revenu supérieures. Locander et Burton (1976) estiment que le degré d'exactitude des données sur le revenu peut dépendre du libellé de la question. Dans une comparaison de quatre libellés, on a constaté que la méthode qui consistait à demander premièrement le montant répète jusqu'à ce que le répondant donne une réponse négative, donnait lieu à une sous-déclaration du revenu. Par contre, l'inversion de l'ordre des paliers du revenu produisait une sur-déclaration du revenu. La question qu'on utilisait pour les interviews téléphoniques dans les études du SCPR était semblable à la première méthode. C'est celle

Tableau 7

Revenu brut des ménages: études du SCPR

Etude n° 1		Etude n° 2	
Par téléphone (183)		Par téléphone (297)	
Sur place (170)		Sur place (352)	
Bases ^a	Revenu		
Inférieur à £5,000	38	27	28
£5,000-£9,999	42	37	37
£10,000 +	21	35	35
Etude n° 1: $\chi^2 = 10.08$ d.l. = 2 $p < 0.01$			
Etude n° 2: $\chi^2 = 0.11$ d.l. = 2 $p > 0.9$			
Bases	(217)	(199)	(344)
N'a pas répondu	16%	15%	14%
N'a pas répondu	13%		
^a "Ne sait pas" et "N'a pas répondu" sont exclues.			

Les enquêtes expérimentales du SCPR ne comportaient pas de question ouverte, mais l'enquête du MRDF renfermait un certain nombre de questions visant à vérifier si la compréhension était spontanée. La comparaison des résultats des interviews téléphoniques et des interviews sur place semble confirmer les observations ci-dessus. À titre d'exemple, le tableau 6 montre qu'un tiers des personnes interviewées par téléphone n'ont pas donné de réponse à une question particulière tandis que dans le cas des personnes interviewées sur place, la proportion correspondante a été inférieure à 25%. En outre, le nombre moyen de réponses données par téléphone a été beaucoup moins élevé que le nombre moyen de réponses données dans une interview sur place. Nous pourrions supposer que des réponses plus nombreuses ou plus longues produisent des résultats plus justes, ce qui nous amènerait à souligner la nécessité de méthodes permettant de mieux adapter les questions ouvertes aux enquêtes par téléphone. À l'opposé, nous pourrions conclure que les questions ouvertes n'ont qu'un usage limité dans les enquêtes par téléphone, notamment lorsqu'on cherche à obtenir uniquement la première impression des répondants. Cette affirmation doit toutefois être vérifiée; nous devons ici nous contenter d'en décrire les effets.

On a également observé des différences de distribution de réponses pour des questions qui avaient trait à des échelles d'attitude et qui ont été incluses aussi bien dans des interviews sur place que dans des interviews téléphoniques. Les personnes interviewées par téléphone tendent à souscrire à ce qu'on leur propose ou à fournir des réponses extrêmes (Jordan, Marcus et Reeder 1980; Groves et Kahn 1979). Le modèle de réponse ("d'accord/pas d'accord") utilisé dans l'enquête du MRDF a permis de constater que les personnes interviewées par téléphone étaient légèrement plus portées à être d'accord avec ce qu'on leur proposait. Par ailleurs, on n'a observé aucune variation notable de l'éventail de réponses; rien ne laissait croire à une plus forte tendance à fournir des réponses extrêmes.

3.1.3 Questions délicates

En ce qui a trait aux genres de questions que l'on peut utiliser dans les interviews téléphoniques, les spécialistes se sont intéressés longuement aux questions délicates, c'est-à-dire à celles qui ont trait à des renseignements personnels et à celles pour lesquelles il existe des réponses plus conformistes que d'autres au point de vue social. Au départ, on ne s'entendait pas sur les effets probables de questions délicates dans une interview téléphonique. D'une part, on prétendait que les répondants seraient moins portés à fournir des réponses justes parce que, l'intervieweur n'étant pas sur place, ils se sentiraient moins obligés d'adopter une attitude franche et ouverte. D'autre part, on prétendait que les répondants seraient plus portés à donner des réponses justes parce que, grâce au caractère impersonnel de la conversation téléphonique, ils hésiteraient moins à divulguer des renseignements à caractère confidentiel. La plupart des constatations tendent à corroborer la seconde thèse (Colombotos 1965; Wiseman 1972; Henson, Roth et Cannell 1974; Locander 1974; Rogers 1976). La principale exception est signalée par Groves et Kahn (1979), qui ont constaté que certaines personnes interviewées par téléphone hésitaient à fournir des renseignements sur leur situation financière et à répondre à d'autres questions délicates. Nos études confirment l'hypothèse selon laquelle les personnes interviewées par téléphone ne craignent pas de répondre honnêtement aux questions délicates. Dans notre première étude par exemple, nous avons déterminé 14 questions qui pouvaient être jugées délicates et les avons soumises à des tests visant à comparer les effets du mode d'interview. Nous donnons ci-dessous trois exemples de questions jugées délicates:

i) Comment qualifieriez-vous votre attitude à l'égard des gens d'autres races?

(Lire à haute voix) ...

... Très empreinte de préjugés

... Légèrement empreinte de préjugés

... Libre de préjugés.

Tableau 5

Durée des interviews selon le mode d'interview (étude n° 2)

Minutes	Bases non pondérées	
	Interview téléphonique (354)	Interview sur place (360)
Moins de 20	10	5
20-29	63	53
30-40	22	33
40 +	6	8

$$\chi^2 = 17.6 \text{ d.l.} = 3 \text{ p} < 0.01$$

Tableau 6

Réponses à une question ouverte (enquête du MRDP):

comparaison par mode d'interview

Que trouvez-vous de bon à la soupe . . . ?

Bases	Nombre de réponses	
	Interview téléphonique (700)	Interview sur place (601)
Aucune	33	22
Une	58	61
Deux	7	14
Trois ou plus	1	2
Moyenne	0.77	0.96

$$\chi^2 = 32.2 \text{ d.l.} = 3 \text{ p} < 0.01$$

L'interview téléphonique, comme le soulignent, par exemple, Dillman (1970) et Williams (1977). Dans une conversation téléphonique, l'interv intervieweur et le répondant ont tous deux tendance à parler plus rapidement et à éviter les pauses. Notre seconde étude a mis en évidence cette constatation. Le tableau 5 donne la durée réelle d'interviews qui, normalement, devaient s'étendre sur 25 minutes. Nous voyons que 10% des interviews téléphoniques ont été réalisées en moins de 20 minutes, comparativement à 5% pour les interviews sur place. À l'autre extrême, 41% des interviews sur place ont duré plus d'une demi-heure, comparativement à moins du tiers pour les interviews téléphoniques.

Ball (1980) estime que l'interview téléphonique suit un rythme plus rapide parce que les règles de la conversation téléphonique exigent que l'intervieweur et le répondant s'efforcent l'un et l'autre de soutenir la conversation. Cela laisse probablement moins de temps aux répondants pour préparer leurs réponses. Sans doute, les moments de silence semblent mettre les gens mal à l'aise; dans une étude de Jordan (1980), les intervieweurs ont qualifié d'interviewables les pauses qu'ils faisaient normalement dans leurs interviews. Il y a sans doute beaucoup d'autres facteurs qui entrent en ligne de compte; même l'absence de signaux visuels n'est pas à négliger.

Tableau 4
Différences entre les distributions de réponses marginales:
interview téléphonique et interview sur place

Bases			
Etude n° 1	(95)	Etude n° 2	(69)
%		%	
Aucune différence significative		87	
Différence significative		9	
à un seuil de 5%		7	
Différence significative		4	
à un seuil de 1%		2	

des distributions ayant traité à des données qui ont été pondérées en fonction des différences qui pouvaient exister entre le nombre de personnes inscrites à une même adresse sur la liste électorale et le nombre de personnes demeurant effectivement à cette adresse. De telles différences ont été relevées dans environ 25% des cas, et chaque fois les données ont été pondérées par le nombre de personnes âgées de 18 ans ou plus qui vivaient à l'adresse divisé par le nombre d'électeurs inscrits à cette adresse. Nous avons choisi de produire des tableaux de données pondérées au cas où le lecteur voudrait tirer des conclusions distinctes pour les données obtenues par interview téléphonique et pour celles obtenues par interview sur place en considérant que les deux séries de données ont été établies selon les procédures normales. On a appliqué des tests de chi-deux standard même si les données provenaient d'un sondage à plusieurs degrés. Il a été démontré (voir, par exemple, Holt, Scott et Ewings 1980) que si l'on sous-estimait la variabilité réelle en ne tenant pas compte du plan de sondage, on obtiendrait en général des variables à tester trop élevées et, par conséquent, on rejetterait à tort des hypothèses nulles (les tests seraient anti-conservatifs). Pour ce qui a trait à l'enquête sur les attitudes sociales, toutefois, l'estimation des erreurs types réelles pour les variables d'attitude produit des effets du plan de sondage (rapport de l'erreur type complexe à l'erreur type de l'échantillonnage aléatoire simple) qui dépassent rarement 1.2 (Jowell et Witherspoon 1985). On soutient en outre dans les ouvrages portant sur le sujet que la répartition en grappes est susceptible d'avoir des conséquences moins sérieuses dans les tests d'indépendance bilatéraux (Holt, Scott et Ewings 1980). C'est pourquoi nous croyons justifiés d'utiliser des tests du chi-deux standard afin d'éviter les nombreux calculs requis pour obtenir des statistiques corrigées. Cette méthode aura plutôt pour effet de surestimer la signification des différences entre les modes d'interview.

Nous avons analysé 95 questions et sous-questions dans la première étude et 69 dans la seconde. Les résultats figurent dans le tableau 4.

Nous voyons clairement que les résultats des deux études sont conformes à ceux obtenus par d'autres spécialistes; les différences de distribution de réponses entre les modes d'interview ne sont significatives que pour un très faible pourcentage des questions. L'étude du MRDF a abouti à une conclusion semblable.

3.1.2 Comparaison de formes de questions particulières

Malgré les constatations générales, des recherches aux Etats-Unis ont montré qu'il existe certains types de questions qui entraînent inévitablement des différences de distributions de réponses significatives. Par exemple, Groves et Kahn (1979) ont démontré que les répondants avaient tendance à donner des réponses incomplètes à des questions ouvertes au cours d'interviews téléphoniques. Cette tendance peut être attribuable au rythme plus rapide de

Pour connaître les raisons qui incitent certains répondants à refuser une interview téléphonique- que, on a approché 55 personnes qui avaient refusé une telle interview lors de la première enquête pour savoir si elles auraient accepté de participer à l'enquête au premier contact si ce contact s'était fait en personne plutôt qu'au téléphone. Quarante de ces personnes ont indiqué que le mode d'interview n'influencerait aucunement leur décision et quelques-unes seulement de ces personnes ont accepté par la suite de se prêter à une interview. Les quinze autres personnes ont pour la plupart affirmé qu'elles auraient participé à l'enquête s'il s'était agi d'une interview sur place, et se sont effectivement prêtées à une telle interview ultérieurement (13 personnes sur 15).

Comme on n'a pas approché les personnes qui avaient refusé une interview sur place, nous ne savons pas si une proportion d'entre elles auraient préféré être interviewées par téléphone.

3.1 Divergence des réponses et qualité des données

La conception que le public se fait d'une utilisation normale du téléphone peut influencer non seulement sur les taux de réponse mais aussi sur les sortes de questions auxquelles les répondants sont disposés à répondre. Ce qui importe davantage, toutefois, c'est le genre de communication qu'il est possible d'établir entre l'intervieweur et le répondant et les effets qu'il peut avoir sur les estimations.

L'interview sur place donne lieu à des échanges verbaux et non verbaux tandis que l'interview téléphonique n'a qu'une capacité d'information limitée, les échanges entre l'intervieweur et le répondant se limitant aux propos de la conversation et à ce qu'on appelle les signaux paralinguistiques: ton de voix, pauses et ainsi de suite (Miller et Cannell 1982).

La capacité d'information limitée de l'interview téléphonique peut avoir de nombreux effets sur les estimations d'enquête. Par exemple, l'absence de support visuel peut rendre plus difficile de répondre à certaines questions. Une communication uniquement verbale peut ne pas rendre entièrement le sens des paroles d'un répondant (ce qui complique, par exemple, l'analyse des réponses à des questions ouvertes) et ne permet pas toujours de vérifier si le répondant a bien compris la question. L'absence de support visuel peut également nuire à l'intervieweur dans l'accomplissement de sa tâche. Par exemple, les signaux verbaux peuvent-il remplacer les signaux non verbaux qui soutiennent l'intérêt et l'attention du répondant ou qui aident à diriger l'interview? L'intervieweur peut-il garder constamment l'attention du répondant, surtout dans les longues interviews? Inversement, est-il souhaitable de supprimer les stimuli visuels afin de réduire les sources de variabilité des données d'enquête? Enfin, le caractère impersonnel de l'interview téléphonique ne fait-il pas hésiter le répondant à fournir des renseignements personnels comme le niveau de revenu ou des renseignements fortement associés à une catégorisation sociale?

Les enquêtes expérimentales du SCPR avaient pour but de répondre à certaines de ces questions.

3.1.1 Comparaison d'ordre général

Compte tenu de la différence de taux de refus entre les deux modes d'interview, il est surprenant de constater que ces modes diffèrent peu à d'autres points de vue. Les mêmes observations ont été faites dans de nombreuses études aux Etats-Unis (Groves et Kahn 1979; Lucas et Adams 1977; Jordan et coll. 1980; Colombotos 1969; Wiseman, 1972) et dans d'autres pays comme le Danemark (Kormendi et coll. 1986). Après avoir utilisé le même questionnaire (questions simples) dans des interviews téléphoniques et des interviews sur place, on a obtenu des distributions de réponses similaires.

Dans les enquêtes du SCPR, on a comparé les distributions de réponses marginales produites par les deux modes d'interview et vérifié à l'aide de tests du chi-deux si les différences observées étaient statistiquement significatives. Ces tests ont été appliqués à des données non pondérées. Toutefois, les tableaux qui figurent dans le présent article contiennent, sauf avis contraire,

Tableau 2
Enquêtes expérimentales du SCPR: effets de la durée de l'interview (étude n° 1)

40 minutes		20 minutes	
Bases		(206)	
		(223)	
		%	
Interviews complètes		48	
Refus		27	
Autres		25	
		18	
		$\chi^2 = 4.7$ d.l. = 2 0.10 > p > 0.05	

Tableau 3
Enquêtes expérimentales du SCPR: effet des lettres d'introduction sur les taux de réponse

Etude n° 1		Etude n° 2	
Lettre		Aucune	
envoyée		lettre	
(215)		(214)	
		%	
Interviews complètes		55	
Refus		23	
Autres		22	
		21	
		15	
		19	
		$\chi^2 = 1.09$ d.l. = 2 p > 0.5	
		$\chi^2 = 2.8$ d.l. = 2 p > 0.2	
		$\chi^2 = 3.49$ d.l. = 2 p > 0.1	

de 40 minutes est principalement attribuable au taux de refus direct plus élevé enregistré pour cette même interview, ce qui donne à penser que les répondants étaient moins disposés à participer à de longues interviews. Néanmoins, seules quelques-unes des personnes qui ont accepté de participer à l'enquête n'ont pas complété l'interview (longue ou courte). De longs questionnaires peuvent exiger une approche différente. Il semble raisonnable de demander à un répondant, au premier contact, de se prêter à une interview de 20 minutes, mais l'intervieweur aurait intérêt à fixer des rendez-vous pour les interviews plus longues. Wiseman et McDonald (1979) estiment que les taux de refus seront moins élevés si les intervieweurs ont reçu l'instruction de fixer une date et une heure précises avec le répondant pour un second contact si leur interlocuteur leur indique qu'il est occupé au moment du premier contact. Dans d'autres études, on a observé que l'envoi de lettres d'introduction à des personnes choisies pour participer à une interview téléphonique avait pour effet d'accroître les taux de réponse. Par exemple, Dillman, Gallegos et Frey (1976) ont constaté des taux de refus de 6% en moyenne parmi les personnes qui avaient reçu une lettre d'introduction, comparative-ment à des taux de 14% pour les personnes qui n'avaient pas reçu une telle lettre. En ce qui concerne les enquêtes du SCPR, le tableau 3 montre des taux de réponse légèrement plus élevés pour les répondants à qui on avait fait parvenir une lettre d'introduction (on ignore toutefois quelle proportion de ces répondants ont vraiment reçu la lettre); néanmoins, les différences observées ne sont pas statistiquement significatives.

méthodes d'enquête.

Le tableau 1 donne les taux de réponse pour les deux enquêtes réalisées par le Centre des

Les taux de réponse des deux enquêtes ont été relativement faibles tant pour les interviews téléphoniques que pour les interviews sur place. (Les taux de réponse pour les interviews sur place devaient normalement être supérieurs à 70% avant la relance des cas de refus). Ces taux anormalement faibles ne sont pas sans rapport avec la nature des enquêtes ; on sait combien il est difficile de convaincre d'éventuels répondants de participer à des enquêtes à caractère général. La même observation vaut pour la seule autre enquête comparative d'importance au Royaume-Uni, réalisée par Marplan pour le compte du Market Research Development Fund. Pour cette enquête, on a utilisé la même méthode d'échantillonnage que celle dont nous nous sommes servis dans nos enquêtes expérimentales et on a posé toute une série de questions d'ordre général sous le titre *Mode de vie des années 80*. Les taux de réponses ont été de 45% pour les interviews téléphoniques et de 67% pour les interviews sur place (Market Research Development Fund, 1985). Les tailles d'échantillon étaient de 1697 pour les interviews téléphonique et de 1233 pour les interviews sur place. Pour ce qui a trait à nos deux enquêtes expérimentales, les taux de réponse ont été moins élevés dans le cas des interviews téléphoniques. À peine la moitié de l'échantillon a produit des interviews complètes. Comme le tableau 1 le montre, la différence entre les taux de réponse était presque significative pour l'étude n° 1, et significative pour l'étude n° 2, et la combinaison les deux études.

On pourrait attribuer les différences de taux de réponse entre les deux modes d'interview à notre inexpérience en matière d'interview téléphonique ; mais ces différences sont également observées dans d'autres pays. Aux États-Unis par exemple, un certain nombre d'auteurs (par exemple Hochstim 1967 ; Henson, Roth et Cannell 1977) ont signalé des taux de réponse moins élevés pour les interviews téléphoniques, cette situation étant surtout attribuable à l'augmentation du nombre de cas de refus. Groves et Kahn résument la situation en ces termes : "Le taux de réponse des enquêtes nationales demeure au moins à cinq pourcent au-dessous du taux de réponse normalement obtenu dans les interviews sur place. Ce résultat est plutôt stable en dépit des changements survenus au cours des années dans la formation des intervieweurs et du perfectionnement des méthodes de supervision, du processus de rétroaction avec les superviseurs et des méthodes de présentation de l'enquête." (Traduction) (Groves et Kahn 1979, p. 219).

Ces constatations donnent à penser que le mauvais accueil réservé à l'interview téléphonique s'explique peut-être plus par des considérations sociologiques et psychologiques que par l'inexpérience de l'intervieweur ou la nouveauté des méthodes utilisées. Cependant, la première enquête du SPCR semble avoir été mieux réussie que la seconde enquête ou l'enquête du MRDF. Certains attribuent cela à l'intérêt et à l'enthousiasme qu'a soulevé la première enquête. Cet enthousiasme peut s'être communiqué aux intervieweurs (par exemple les spécialistes assistaient souvent aux interviews), ce qui aurait influé sur le taux de réussite de l'enquête. Les résultats obtenus dans le cas des interviews sur place donnent certes à penser que l'enthousiasme et le dynamisme de l'intervieweur sont des éléments essentiels pour obtenir des taux de réponse élevés.

Dans les enquêtes du SPCR, on a modifié deux conditions pour évaluer l'effet que cela aurait sur les taux de réponse pour les interviews téléphoniques. En ce qui a trait à la première enquête, la moitié des personnes interviewées par téléphone se sont prêtées à une interview de 20 minutes et l'autre moitié à une interview de 40 minutes (on informait les répondants de la durée de l'interview vers la fin de la présentation) ; par ailleurs, pour les deux enquêtes, on a fait parvenir à une moitié de l'échantillon de personnes interviewées par téléphone (cette moitié étant choisie au hasard) une lettre leur annonçant l'interview.

Le tableau 2 montre que le taux de réponse est inférieur dans le cas de l'interview de 40 minutes mais qu'il n'y a pas de différence significative entre les distributions (en pourcentage) pour les deux genres d'interview. Le plus faible taux de réponse enregistré pour l'interview

Il convient de se rappeler toutes ces contraintes au moment de l'étude des résultats; néanmoins, ces contraintes sont presque inévitables dans des études comparatives. Comme nous l'avons vu ci-dessus, nous avons tenté de définir et de minimiser ces limites. Celles-ci sont préoccupantes seulement quand les résultats de nos enquêtes donnent à penser que des effets de modes d'interview pourraient brouiller les effets d'autres variables; or, la plupart de nos résultats ne laissent entrevoir rien de tel. Nous pourrions, par conséquent, considérer que les limites définies ci-dessus ne peuvent avoir que des effets qui neutraliseraient les effets de modes d'interview que nous observions en d'autres circonstances; de sorte que la validité de nos conclusions est menacée dans une moindre mesure.

3. TAUX DE RÉPONSE

Au Royaume-Uni, on s'interroge beaucoup sur la faisabilité de l'interview téléphonique, surtout pour des enquêtes sociales. On se demande non seulement quel niveau de communication est possible et quel pourra en être l'effet sur les aspects objectifs et affectifs de l'interview, mais encore s'il est socialement acceptable d'utiliser le téléphone pour faire des enquêtes. En Grande-Bretagne, les spécialistes des sciences sociales croient généralement que les appels téléphoniques d'étrangers sont susceptibles d'être accueillis avec méfiance; un appel d'un intervieweur pourrait être considéré comme inopportun et gênant.

En contrepartie, on souligne souvent les avantages que peut présenter l'interview téléphonique par rapport à l'interview sur place, notamment dans les quartiers situés au coeur des grandes villes. L'intensification des crimes contre la personne et la propriété a amené les gens à se méfier de plus en plus des étrangers, attitude qui se traduit par une diminution des taux de réponse; en outre, l'installation de dispositifs comme des systèmes d'intercommunication rend plus difficile le contact avec des répondants éventuels. En revanche, si l'intervieweur utilise le téléphone, il pourra sans doute rejoindre un répondant éventuel s'il y a quelque'un à la maison et, dans le cas contraire, il ne sera pas beaucoup plus coûteux de rappeler plus tard.

Tableau 1
Enquêtes expérimentales du SCPR: taux de réponse

	Etude n° 1		Etude n° 2	
	Par téléphone	Sur place	Par téléphone	Sur place
Bases	(429)	(313)	(730)	(631)

Interviews complètes	53	60	46	68
Interviews partielles	1	-	-	-
Refus (sans sélection)	5	2	21	6
Refus (personne interposée)	9	5	7	4
Refus (personne échantillonnée)	11	18	10	11
Aucun contact ^a	3	1	8	4
Personne échantillonnée toujours absente	3	3	3	2
Malade, à l'extérieur - ne comprends pas la langue de l'intervieweur	2	5	2	4
Autre ^b	13	6	4	2

Etude n° 1: $\chi^2 = 3.72$ d.l. = 1 0.05 < p < 0.1
Etude n° 2: $\chi^2 = 66.22$ d.l. = 1 p < 0.001
Etude n° 1 et 2 combinées: $\chi^2 = 59.46$ d.e. = 1 p < 0.005

Comparaisons utilisant deux catégories
seulement:
interviews complètes et
interviews non complètes

^a Incluant "Pas de réponse" et "Ligne constamment occupée".
^b Incluant "Rendez-vous manqué", "Trop âgé", "Handicapé", "Pas de service au numéro composé", "Bon numéro, mauvaise adresse".

à penser qu'au-delà des différences de taux de réponse globaux entre les modes d'interview, certains types de personnes sont plus susceptibles de se prêter à une interview téléphonique qu'à une interview sur place et vice-versa. Les variables utilisées sont l'âge croisé avec le sexe, l'état matrimonial, la composition du ménage, la situation économique, le groupe socio-économique et le lieu de résidence. La première étude n'a révélé aucune différence significative sur le plan statistique entre les taux de non-réponse. Dans la seconde étude, des différences significatives entre l'échantillon interviewé par téléphone et l'échantillon interviewé sur place ont été observées pour deux variables: la composition du ménage (l'échantillon de personnes interviewées par téléphone comportait relativement plus de couples de moins de 60 ans sans enfant, tandis que l'échantillon de personnes interviewées sur place comportait relativement plus de couples avec des enfants en bas âge ou des adolescents), et le groupe socio-économique (les cols blancs de niveau subalterne ou intermédiaire et les travailleurs ou travailleurs des "autres" professions étaient plus fortement représentés dans l'échantillon interviewé par téléphone que dans l'échantillon interviewé sur place, tandis que les "personnes au foyer" étaient relativement plus nombreuses dans l'échantillon interviewé sur place). Ces différences ne représentent peut-être que des fluctuations d'échantillonnage mais elles devraient inspirer une certaine prudence dans l'interprétation des différences dans les réponses des deux échantillons. Le deuxième facteur qui pourrait limiter la comparaison des deux modes d'interview sont les différences de niveau de compétence ou de supervision entre les intervieweurs affectés aux interviews téléphoniques et ceux affectés aux interviews sur place. Six intervieweurs du premier groupe ont participé à la première enquête expérimentale. Deux de ces six personnes avaient de l'expérience dans les interviews sur place tandis que les quatre autres n'avaient aucune expérience; ces dernières ont donc reçu une formation de base ainsi que la formation spéciale pour les interviews téléphoniques, qui a aussi été offerte aux deux intervieweurs expérimentés. La seconde enquête réunissait dix intervieweurs, dont trois avaient participé à l'enquête précédente. Comme dans le premier cas, un superviseur était présent pour suivre la conversation téléphonique, proposer, si nécessaire, des techniques d'interview et relever les erreurs évidentes dans les questionnaires remplis.

Pour les deux enquêtes, les intervieweurs affectés aux interviews sur place ont été choisis parmi les quelques 300 intervieweurs réguliers du Social Community Planning Research. Ces personnes ont reçu la même formation de base que les personnes affectées aux interviews téléphoniques, qui avaient généralement moins d'expérience que celles-ci. Ces différences et particulièrement celles qui font entrevoir une qualité inférieure dans les interviews téléphoniques, devraient être prises en considération.

Le troisième facteur concerne le questionnaire. Le questionnaire de l'enquête principale, qui comprenait environ 100 questions, couvrait cinq grands sujets: l'emploi, l'éducation, la santé et le logement, les classes sociales et la discrimination fondée sur la race ou le sexe. Les questionnaires des deux enquêtes expérimentales étaient composés des questions jugées les plus importantes dans l'enquête principale. Ces questions ont été choisies de manière à refléter tout l'éventail des questions de l'enquête principale.

Les questionnaires des enquêtes expérimentales couvraient donc toute une série de sujets (y compris certains sujets "délicats") et renfermaient des questions qui correspondaient à divers modèles de réponse et niveaux de complexité. Les questions utilisées dans les interviews de 20 et de 40 minutes de la première enquête et dans l'interview de 25 minutes de la seconde enquête ont été posées dans le même ordre qu'elles l'avaient été dans l'enquête principale. Ainsi, l'interview de 40 minutes ne consistait pas en deux interviews de 20 minutes chacune (la première d'entre elles étant la version existante) mais les questions d'une interview de 20 minutes étaient modifiées uniquement lorsque cela était vraiment nécessaire; par exemple, certaines questions ont dû être reformulées pour tenir compte de l'absence inévitable de cartons ou de fiches. Comme le questionnaire de l'enquête sur les attitudes sociales est surtout constitué de questions fermées, quelques-uns seulement des résultats de nos enquêtes expérimentales ont trait à des questions ouvertes.

Les numéros de téléphone ainsi obtenus ont été répartis systématiquement en quatre sous-échantillons. Les membres de deux de ces sous-échantillons ont été interviewés par téléphone à l'aide d'un questionnaire qui devait être rempli en l'espace d'environ 20 minutes. Les questions étaient tirées de toutes les sections du questionnaire de l'enquête sur les attitudes sociales. Les membres des deux autres sous-échantillons ont été interviewés par téléphone à l'aide d'un questionnaire plus long (qui devait être rempli en 40 minutes environ), dont les questions provenaient également du questionnaire de l'enquête principale. Les membres des quatre sous-échantillons ont reçu une lettre les avisant qu'ils allaient recevoir un appel téléphonique pour une enquête. Les membres des autres sous-échantillons n'ont reçu aucun avis à cet effet. Dans tous les cas, le choix des répondants s'est fait selon les mêmes règles que celles utilisées pour l'enquête principale.

Les membres de l'échantillon expérimental pour lesquels il n'a pas été possible d'obtenir un numéro de téléphone ont été soumis à une interview sur place de 20 minutes. Les résultats des interviews téléphoniques et des interviews sur place (méthode d'interview mixte) ont ensuite été comparés aux résultats de l'enquête principale, qui comportait uniquement des interviews sur place (Sykes et Hoinville, 1985).

On a voulu analyser plus directement les effets du mode d'interview en soumettant au service d'extraction de numéros de téléphone de British Telecom un sous-échantillon systématique de 600 adresses choisies parmi les adresses utilisées pour l'enquête principale (5 adresses par point d'échantillonnage). British Telecom a alors produit des numéros de téléphone pour 55% des adresses qui lui avaient été soumises (la variabilité du taux d'extraction demeure inexpliquée). On a ensuite comparé les réponses fournies par les personnes interviewées par téléphone et celles fournies par les personnes qui auraient pu être interviewées par téléphone mais l'ont été sur place. En limitant la comparaison aux personnes qui peuvent être rejointes par téléphone, nous avons éliminé les effets attribuables aux différences de méthode de collecte de données et avons considéré uniquement les effets découlant des différences entre les groupes comparés.

2.2 Etude n° 2

La seconde étude portait plus spécialement sur cette comparaison directe. On a prélevé environ 2300 adresses dans la liste électorale, comme dans la première étude, et on les a soumises à British Telecom pour l'extraction des numéros de téléphone (dans ce cas, le taux d'extraction a été de 61%). Les adresses pour lesquelles British Telecom a produit des numéros de téléphone ont été réparties en trois sous-échantillons. Les membres du premier sous-échantillon ont été interviewés par téléphone suivant des méthodes classiques; les membres du deuxième sous-échantillon ont été interviewés à l'aide de la méthode ITAO (interview téléphonique assistée par ordinateur) et les membres du troisième sous-échantillon ont été interviewés sur place. L'application de la méthode ITAO a été un échec (pour diverses raisons), mais les deux autres sous-échantillons nous permettent, encore une fois, de faire une comparaison directe entre les personnes interviewées par téléphone et celles qui auraient pu l'être mais qui ont été interviewées sur place. Le questionnaire, qui devait être rempli en 25 minutes, formait un sous-ensemble de questions de l'enquête de 1983 sur les attitudes sociales au Royaume-Uni.

2.3 Limites de la comparaison des modes d'interview

Trois facteurs peuvent limiter la comparaison entre les réponses obtenues par l'interview sur place et celles obtenues par l'interview téléphonique. Premièrement, des différences de taux de non-réponse (voir section 3) peuvent avoir été à l'origine de différences dans la composition des groupes de répondants. Cette possibilité a été vérifiée à l'aide d'un certain nombre de variables démographiques et socio-économiques que l'on croit être associées à certaines variables d'attitude. Des différences significatives entre les groupes de répondants donnent

Deux comparaisons ont été faites dans la première étude: la première entre un échantillon expérimental soumis à l'interview mixte et un échantillon national, plus grand, soumis à l'interview sur place; la seconde entre deux échantillons de personnes qui ont le téléphone, le premier échantillon étant interviewé par téléphone et l'autre sur place. Notre étude porte plus spécialement sur la seconde comparaison, qui soulève la question qui est à la base de toute évaluation de la méthode d'interview mixte: les données obtenues par l'interview téléphonique et l'interview sur place sont-elles compatibles ou y a-t-il des divergences du fait de l'utilisation de deux modes d'interview différents? Dans ce dernier cas, il est impossible de "d'ajouter" les données et de les considérer comme une seule série sans faire certains ajustements qui, normalement, ne sont pas possibles dans le cadre d'une enquête unique. La seconde étude porte uniquement sur la comparaison directe des deux modes d'interview auxquels sont soumises les personnes qui ont le téléphone.

2.1 Etude n° 1

La première étude a été réalisée concurrentement avec l'enquête de 1983 sur les attitudes sociales au Royaume-Uni (1983 British Social Attitudes Survey), qui est désignée ici comme l'enquête "principale". Cette enquête comportait des interviews sur place d'une durée approximative d'une heure, qui couvraient un large éventail de sujets d'ordre politique, économique, social et moral.

L'enquête principale était formée d'environ 1750 répondants âgés de 18 ans ou plus et membres de ménages privés. Pour des raisons pratiques, l'échantillon était limité aux personnes dont le nom figurait sur la liste électorale. Les pensionnaires d'institution (à l'exception des ménages privés vivant en institution) étaient exclus du champ de l'enquête, comme l'étaient les 4% d'adultes dont l'adresse de résidence ne figurait pas sur la liste électorale. (Todd et Butcher 1982).

Pour l'enquête, on a utilisé un plan de sondage à plusieurs degrés comportant quatre étapes de sélection: tout d'abord, 103 circonscriptions d'Angleterre et du pays de Galles et 11 districts municipaux d'Ecosse ont été sélectionnés avec une probabilité proportionnelle à la taille de l'électorat; ensuite, une section de vote a été sélectionnée dans chacun de ces points d'échantillonnage avec, là encore, une probabilité proportionnelle à la taille de l'électorat; en troisième lieu, vingt-trois adresses ont été choisies dans chacune des sections de vote avec une probabilité proportionnelle au nombre d'électeurs inscrits à l'adresse. Enfin, l'intervieweur a choisi une personne à chaque adresse échantillonnée à l'aide d'une variante de la méthode de Marchant-Blyth. (Blyth et Marchant 1973).

Pour les besoins de l'étude, on a prélevé un échantillon parallèle d'environ 800 adresses des 114 points d'échantillonnages originaux (7 adresses par point d'échantillonnage). On a ensuite soumis ces adresses et les noms inscrits sur la liste électorale au service d'extraction des numéros de téléphone de British Telecom. Le service a produit des numéros de téléphone pour 65% des adresses qui lui avaient été soumises. La différence entre ce taux d'extraction et la proportion de ménages ayant le téléphone – environ 75% au moment de l'étude – s'explique principalement par les numéros qui ne sont pas inscrits dans l'annuaire; en effet, selon Collins et Sykes (1987), environ 12% des numéros de téléphone en Grande-Bretagne ne figurent pas dans l'annuaire, la proportion variant selon la région et d'autres facteurs. Les autres problèmes liés à l'extraction des numéros de téléphone semblent avoir eu peu d'effet sur la bonne marche de l'étude.

British Telecom a procédé de la façon suivante pour extraire les numéros de téléphone: après avoir déterminé la circonscription téléphonique appropriée au moyen de l'adresse, on a cherché le nom de l'abonné dans l'annuaire. Les détails de l'adresse (par exemple le nom de la rue) permettaient de distinguer les abonnés qui avaient des noms identiques. Comme la liste électorale n'indiquait pas quelle personne parmi celles demeurant à la même adresse était l'abonné du téléphone, British Telecom avait pour instruction de vérifier chaque nom avant de rejeter l'unité d'échantillonnage.

Comparaison entre l'interview téléphonique et l'interview sur place au Royaume-Uni

W.M. SYKES et M. COLLINS¹

RÉSUMÉ

Dans cette étude, les auteurs présentent les résultats des expériences de méthodologie qui ont servi à comparer l'interview téléphonique et l'interview sur place dans les enquêtes menées auprès de la population en général. Le nombre relativement élevé de personnes qui n'ont pas le téléphone au Royaume-Uni, surtout chez les plus défavorisées, est un argument qui milite en faveur de l'application d'une méthode d'interview mixte qui combinerait l'interview téléphonique et l'interview sur place, pour ceux ou celles qui n'ont pas le téléphone. L'efficacité de cette méthode repose sur l'absence, dans les réponses d'enquête, de distorsions attribuables au caractère mixte de la méthode ou sur la possibilité de tenir compte de telles distorsions s'il y en a.

MOTS CLÉS: Interview téléphonique; interview mixte; enquêtes sociales; taux de réponse; qualité des données.

1. INTRODUCTION

Le choix d'une méthode de collecte de données pour une enquête dépend des renseignements dont on dispose sur les diverses méthodes possibles. Au Royaume-Uni, l'interview téléphonique est une méthode qui commence à peine à être connue. Ce n'est que depuis deux ans que l'on compare l'interview téléphonique aux autres méthodes de collecte de données. On peut s'étonner de cette méthode relativement nouvelle vu le caractère animé du débat sur les avantages et les inconvénients de l'interview téléphonique et l'attention qu'a reçue cette question dans d'autres pays. Le contenu du présent article est fondé principalement sur deux études qui ont été réalisées par le Centre de méthodes d'enquête du Social and Community Planning Research et qui visaient à comparer l'interview téléphonique et l'interview sur place. Réalisées en 1983 et en 1984, ces études traitent quelques-unes des questions fondamentales, à savoir le fait pour le public d'être plus ou moins disposé à participer aux enquêtes par téléphone et le genre, la qualité et la quantité de données qui peuvent être recueillies. Les deux études sont décrites dans la section 2 et leurs résultats présentés dans les sections 3 et 4. Il est également question dans ce document d'une autre étude britannique – une expérience réalisée en 1985 par le Market Research Development Fund – et des activités de recherche méthodologique plus nombreuses qui ont lieu dans d'autres pays, particulièrement les États-Unis.

2. LES ÉTUDES DU SCPR

Nos recherches ont reflété le fait que le pourcentage de ménages qui possédaient un téléphone au Royaume-Uni était faible par rapport à l'Amérique du Nord: en effet, environ 75% des ménages britanniques avaient le téléphone en 1983. La non-exhaustivité est appréciable et elle est déterminante pour les spécialistes des enquêtes sociales parce qu'elle introduit un biais à l'égard des couches moins favorisées de la société britannique. Dans les circonstances, les deux études visaient essentiellement à évaluer la méthode d'interview mixte, selon laquelle les personnes qui ont le téléphone seraient interviewées par téléphone et celles qui ne l'ont pas seraient interviewées sur place.

¹ W.M. Sykes et M. Collins, Survey Methods Centre, SCPR, 35 Northampton Square, Londres, EC1V 0AX, Angleterre.

- KISH, L. (1976). Optima and proxima in linear sample designs. *Journal of the Royal Statistical Society, Sér. A*, 139, 80-95.
- KISH, L. (1965). *Survey Sampling*. New York: John Wiley.
- THORNBERY, O.T., et J.T. MASSEY (1983). Coverage and response in random digit dialed national surveys. *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 654-659.
- WAKSBERG, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 40-46.

de réduire les coûts de filtrage au moment de l'élaboration du plan de sondage. Avec d'autres populations rares formées de groupes dont la concentration géographique est aussi nette que celle de la population noire des Etats-Unis, on peut s'attendre aux résultats à peu près analogues.

Deuxièmement, l'utilisation de règles de rejet fondées sur des critères d'admissibilité des unités d'une sous-population réduit effectivement les coûts de filtrage dans les UPE. Cela augmente la proportion admissible de numéros secondaires de deux à neuf fois selon la densité de la strate considérée.

Troisièmement, on ne perd pas beaucoup de numéros admissibles en utilisant des UPE plus grandes (c'est-à-dire des UPE de 200 plutôt que de 100 numéros consécutifs). Les séries de 100 à forte densité de numéros admissibles ont tendance à précéder ou à suivre d'autres séries à forte densité. Cela est un phénomène que nous n'avions pas encore observé concernant l'attribution des numéros par les compagnies de téléphone. Ce fait permet de prélever plus de numéros par UPE, ce qui est un autre élément clé de la réduction des coûts de production de l'échantillon de la population noire.

Malgré qu'on soit fortement tenté de chercher davantage à réduire les coûts dans le cas des enquêtes par sondage auprès de populations rares, il importe d'établir explicitement un juste compromis entre les erreurs et les coûts au moment de décider du plan de sondage définitif. Le modèle de coûts et d'erreurs d'échantillonnage que nous avons adopté pour notre recherche donne à penser qu'il est préférable de ne pas répartir l'échantillon proportionnellement entre les strates de population noire et qu'il est plus efficace de choisir un nombre relativement grand de numéros secondaires par UPE. Ce nombre devrait être suffisamment grand pour obliger à utiliser des UPE de 200 ou de 400 numéros consécutifs.

Même si nous n'avons appliqué ce plan de sondage qu'à la population noire, les résultats qu'il donne devraient être à peu près les mêmes pour d'autres populations à concentration géographique nette. Ces autres populations incluent notamment les groupes selon le revenu, certaines catégories professionnelles et les groupes ethniques.

En outre, les résultats notre recherche a abouti peuvent également avoir des répercussions sur les méthodes d'échantillonnage transversal. Il se peut qu'il soit avantageux d'augmenter la taille des UPE de 100 à 200 numéros consécutifs dans un plan de sondage par CNH à deux degrés qu'on voudrait appliquer à l'ensemble de la population des ménages abonnés au téléphone. Des séries plus grandes de 200 numéros consécutifs donneraient deux fois plus de numéros parmi lesquels choisir et, comme dans le cas d'enquêtes par sondage auprès de populations rares, la proportion de numéros admissibles aurait tendance à être la même que celle qu'on obtient avec des séries de 100 numéros consécutifs. Par conséquent, avec des valeurs faibles de la corrélation intragrappe, on pourrait décider d'avoir, pour un plan donné, un nombre de numéros admissibles par grappe qui serait beaucoup plus proche de la taille optimale. Parce que toutes les UPE choisies permettraient d'obtenir le nombre choisi de numéros admissibles, la distribution des grappes selon leur taille devrait aussi être moins variable sur l'ensemble des UPE et toute pondération compensatrice devrait donc avoir peu d'effet sur la variance des estimations.

BIBLIOGRAPHIE

- BLAIR, J., et R. CZAJA, (1982). Locating a special population using random digit dialing. *Public Opinion Quarterly*, 46, 585-590.
- GROVES, R.M., An empirical comparison of two telephone sample designs. *Journal of Marketing Research*, 15, 622-631.
- GROVES, R.M. et R.L. KAHN (1979). *Surveys by Telephone*. New York: Academic Press.

Tableau 6

Répartition optimale de l'échantillon entre les strates pour les moyennes globales, compte tenu de la formation de grappes de taille optimale dans chaque strate, pour diverses ratios des écart-types à l'intérieur des strates et diverses valeurs de la corrélation intragrappe

Ratios des écarts-types à l'intérieur des strates (Forte, moyenne et faible densité)		Ratios des fractions de sondage optimales (Forte, moyenne et faible densité)	
$p = .005$			
3 : 2 : 1	1.7 : 1.5 : 1	1 : 1 : 1	.33 : .5 : 1
$p = .01$			
3 : 2 : 1	1.7 : 1.5 : 1	1 : 1 : 1	.33 : .5 : 1
$p = .02$			
3 : 2 : 1	1.8 : 1.5 : 1	1 : 1 : 1	.33 : .5 : 1
5.1 : 2.7 : 1	3 : 2 : 1	1.7 : 1.3 : 1	.6 : .9 : 1

les fractions de sondage optimales sont dans des proportions de 2.5, 1.6 et 1). D'après ces deux résultats, on peut dire que la méthode utilisée dans l'étude de production et qui consiste à prélever proportionnellement plus d'une unités d'une strate que d'une autre a entraîné, pour un coût unitaire donné, une perte de précision par rapport à la méthode correspondant à l'utilisation de fractions de sondage optimales.

6. SOMMAIRE

L'échantillonnage de populations rares oblige le statisticien qui veut s'en servir pour faire des enquêtes à prendre en considération diverses combinaisons de définitions des UPE et des grappes, de méthodes de stratification et de méthodes de mesure de la taille qu'on ne trouve pas habituellement dans les échantillons transversaux. Notre recherche nous a amenés à conclure qu'il est possible de modifier les techniques de sondage habituelles pour augmenter l'efficacité des échantillons à deux degrés pour les enquêtes téléphoniques auprès de la population des ménages noirs abonnés au téléphone. Premièrement, nous avons constaté que même avec une correspondance très approximative entre les limites géographiques des circonscriptions téléphoniques et celles des grandes villes et des États, il a été possible de stratifier l'univers visé par l'enquête en strates qui permettaient de distinguer des groupes de circonscriptions possédant des taux d'admissibilité très différents. La proportion de ménages noirs était plus que deux fois plus élevée dans la strate à forte densité que dans la strate à faible densité. Cela permet

Tableau 5

Paramètres de coûts et nombre optimal d'éléments par grappe selon la strate pour les grappes tirées des séries de 100 et de 200 numéros consécutifs et différentes valeurs de ρ					
Strate et définition des grappes		Taille optimale des grappes		Paramètres de coûts	
		$\rho = .005$	$\rho = .01$	$\rho = .02$	C_{ha}
Strate à forte densité					
100	15.9	11.2	7.9	\$50.81	\$40.11
200	15.9	11.2	7.9		\$39.78
Strate à densité moyenne					
100	22.4	15.8	11.1	\$114.09	\$45.18
200	21.3	15.0	10.6		\$50.00
Strate à faible densité					
100	29.8	21.0	14.8	\$309.98	\$69.52
200	29.9	21.1	14.8		\$69.18

La deuxième décision importante qu'il nous a fallu prendre au sujet du plan de sondage avait trait au choix de la façon de répartir les unités de l'échantillon entre les strates. L'enquête utilisait des fractions de sondage dans des proportions de 3, 2 et 1 de la strate à forte densité à la strate à faible densité. Nous avons cherché à déterminer quelle serait la répartition optimale de l'échantillon entre les strates en supposant que les grappes choisies dans chaque strate était de taille optimale (voir tableau 5). Etant donné une taille fixée des grappes dans chaque strate, b_h , nous posons que la fraction de sondage dans la h -ième strate, f_h , est proportionnelle à $\sqrt{(Deff_h S_h^2) / (C_{ha} / b_h)}$, où $Deff_h$ est l'effet du plan de sondage pour la statistique dans la h -ième strate, S_h^2 est la variance de l'échantillon dans la h -ième strate, C_{ha} correspond aux coûts d'échantillonnage et de filtrage des UPE dans la h -ième strate et b_h est le nombre d'unités d'échantillon par grappe dans la h -ième strate. Le tableau 6 présente les proportions optimales entre les fractions de sondage pour différentes combinaisons de variances dans chacune des trois strates et diverses valeurs de ρ . Le tableau montre que la répartition optimale de l'échantillon entre les strates est relativement peu sensible aux variations de ρ (pour les valeurs de ρ susceptibles d'être utilisées dans le plan de sondage). Quand les deux strates à plus forte densité de ménages noirs ont une variance au moins égale à celle de la strate à faible densité, il est préférable de prélever proportionnellement plus d'unités de sondage dans ces strates. (Cela reflète les coûts beaucoup moins élevés dans ces strates.) Quand l'écart-type est dans des proportions de 1.7, 1.5 et 1 environ entre les trois strates, le meilleur rapport entre les fractions de sondage est de 3, 2 et 1. L'analyse des données obtenues à partir de l'enquête donne à penser que, pour beaucoup de variables, l'écart-type est à peu près dans des proportions de 1, 1 et 1 entre les trois strates. Dans le cas de ces variables, les fractions de sondage, pour être optimales, doivent avoir un rapport de 1.7, 1.4 et 1 pour les tailles optimales des grappes présentées dans le tableau 5. (Pour une taille de grappe de 6.5, soit la taille effectivement utilisée dans chaque strate,

5. CARACTÉRISTIQUES OPTIMALES DU PLAN DE SONDAGE

Les sections précédentes portaient sur l'effet de l'utilisation de différents plans de sondage sur l'efficacité en termes de coûts et sur la variance d'échantillonnage. Dans la plupart des enquêtes par sondage, c'est à l'étape de l'élaboration du plan de sondage qu'on compare les coûts d'enquête et les erreurs pour chercher à déterminer quelles devraient être les caractéristiques "optimales" de l'enquête. Suivant cette façon de procéder, on tente d'élaborer le plan de sondage qui permettra d'obtenir la plus petite variance possible pour un ensemble donné de ressources à consacrer à l'enquête. Vu la nature de la présente étude, la recherche du choix optimal portera avant tout sur les deux caractéristiques suivantes du plan de sondage: a) nombre d'éléments par UPE et b) répartition de l'échantillon entre les trois strates définies selon la densité de la population noire.

Pour déterminer la taille optimale des grappes, nous avons utilisé le modèle de coûts totaux suivant: $C = C_0 + C_a + C_{ab}$, où C_0 représente les coûts fixes, où C_a représente le coût d'échantillonnage et de filtrage par grappe et a le nombre de grappes choisies et où C_b représente le coût d'échantillonnage, de filtrage et d'interview associé à chaque interview obtenue et b le nombre d'interviews obtenues dans chaque grappe. Comme la proportion de ménages noirs varie d'une strate à l'autre, la valeur des paramètres C_a et C_b varie aussi selon la strate (voir tableau 5). La taille optimale des grappes est donnée par $\sqrt{C_a(1-p)/(C_{bp})}$ (Kish 1965). Le tableau 5 présente aussi les résultats des estimations que nous avons faites pour déterminer quelles seraient les tailles optimales des grappes pour les moyennes et les proportions globales et trois valeurs différentes de la corrélation intragrappe, .005, .01 et .02, en utilisant les données sur les coûts tirées de l'enquête de production. (Ces résultats sont comparables à ceux qui ont été obtenus pour les variables d'attitudes politiques et d'intentions de vote dans les enquêtes réelles.) Les estimations des coûts C_a et C_b pour chaque strate figurent également dans le tableau. Le tableau montre que c'est dans la strate à faible densité que la taille optimale des grappes est la plus grande, ce qui reflète le fait que les coûts de filtrage sont élevés dans cette strate. À noter aussi que ces tailles des grappes optimales tendent à être plus grandes que celles qui ont effectivement été utilisées dans l'enquête (soit $b = 6.5$).

À noter encore que la taille optimale des grappes est à peu près la même pour les grappes tirées des UPE de 100 comme de 200 numéros consécutifs et que la perte d'efficacité en termes de coûts par suite de l'utilisation des séries de 200 plutôt que de 100 numéros consécutifs est minime. (Les estimations de la variance d'échantillonnage signifient également que la corrélation intragrappe est à peu près la même, indépendamment du fait que les grappes sont tirées des séries de 100 ou de 200 numéros consécutifs.)

Les valeurs de la taille optimale des grappes figurant dans le tableau 5 excèdent en général la valeur de la taille des grappes que permettent d'obtenir des UPE de 100 numéros consécutifs. Autrement dit, une proportion élevée d'UPE de 100 numéros consécutifs ne contiendrait pas suffisamment de numéros de ménages noirs pour produire des grappes de la taille désignée au second degré. Cette seule raison suffit à faire préférer les séries de 200 numéros consécutifs. Même avec les séries de 200 numéros consécutifs, il n'a pas été possible d'obtenir les grappes de la taille souhaitée au second degré pour toutes les UPE de la strate à faible densité. (Cela donne à penser que la vraie solution au problème du choix de la taille optimale des grappes devrait tenir compte de la capacité des UPE de produire des grappes; par ailleurs, la méthode utilisée ici peut certainement aider à prendre des décisions pratiques au sujet de l'efficacité en termes de coûts, mais elle ne prend pas en considération certaines situations extrêmes.)

Tableau 4

Etude de production
Corrélations intragrappes synthétiques selon la strate et selon que les grappes sont tirées des séries de 100 ou de 200 numéros consécutifs pour sept statistiques choisies

Corrélation intragrappe synthétique*					
Statistique	Strate à forte densité de population noire	Séries de 100		Séries de 200	
		Séries de 100	Séries de 200	Séries de 100	Séries de 200
Proportion très satisfaite de la vie en général	.021	-.002	-.172	-.042	-.238
Proportion qui estime être dans une meilleure situation financière que l'année précédente	.113	.075	.094	.069	.206
Proportion qui votera pour Mondale	.189	.021	.086	-.087	-.436
Proportion qui va à l'église	.013	.017	-.009	-.078	.035
Proportion ayant passé toute sa vie dans la même ville	-.078	.001	.058	.114	.221
Proportion qui a voté aux élections présidentielles de 1980	-.045	-.035	-.101	-.013	.364
Proportion qui pense que Reagan sera élu président	-.045	-.045	-.545	-.078	.124
Moyenne	.024	.005	-.084	-.016	.039

* Ces estimations ne sont pas pondérées.

déterminant de la variance d'échantillonnage soit la pondération non optimale qu'il faut effectuer pour tenir compte du fait que l'échantillon n'est pas réparti de façon proportionnelle, tandis que l'augmentation de la taille des UPE de 100 à 200 numéros consécutifs ne fait perdre que très peu de précision.

Le tableau 4 présente les résultats du calcul des corrélations synthétiques intragrappe par strate pour les sept statistiques de l'enquête utilisées pour estimer l'effet moyen du plan de sondage. Les estimations des corrélations synthétiques intragrappe ont été obtenues à partir de la valeur de l'effet du plan du sondage à l'aide du modèle de Kish où $Rho = (Def - 1) / (b - 1)$ et n'ont pas été pondérées pour qu'il n'y ait pas d'effet de confusion de la pondération sur les estimations synthétiques. Les estimations du tableau ont tendance à être instables à cause du petit nombre de grappes prélevées dans chaque strate, de l'assez faible taille moyenne des grappes une fois les interviews complétées et du coefficient de variation associé. Ces caractéristiques du plan de sondage rendent plus difficile pour nous de tirer des conclusions à propos de l'effet du regroupement par grappes dans les séries de 100 et de 200 numéros consécutifs. En général, les estimations de la corrélation intragrappe sont un peu plus élevées dans les séries de 100 que dans les séries de 200. Nous croyons que cela reflète plus l'instabilité de la corrélation synthétique estimative qu'une différence réelle au niveau de l'effet du regroupement par grappes et que ces estimations ne donnent pas beaucoup d'évidence que la corrélation intragrappe soit différente selon qu'on la calcule pour des séries de 100 ou de 200 numéros consécutifs.

4. CARACTÉRISTIQUES DE LA VARIANCE D'ÉCHANTILLONNAGE

Pour réduire les coûts de l'échantillonnage de ménages noirs par CNH, il est avantageux d'utiliser de grandes grappes de ménages par UPE (c'est-à-dire d'utiliser moins d'UPE pour une taille d'échantillon donnée) et de tirer proportionnellement plus d'UPE des strates composées des circonscriptions où la proportion de ménages noirs abonnées au téléphone varie. Bien que l'utilisation de plus grosses grappes et la répartition non proportionnelle de l'échantillon améliorent l'efficacité en termes de coûts, la précision globale de l'échantillon est diminuée par l'effet de l'utilisation de plus grosses grappes et par les effets du plan de sondage rendus plus marqués par la nécessité d'effectuer une pondération non optimale pour tenir compte des probabilités inégales de sélection des ménages à partir des trois strates selon la densité de la population noire. Kish (1976) a décrit, les effets augmentés du plan de sondage des estimations dans les cas où la pondération n'est pas optimale. Nous montrons dans les paragraphes qui suivent comment l'utilisation de grappes influe les effets du plan de sondage dans la version modifiée de la méthode d'échantillonnage par CNH utilisée dans la présente étude.

Toutes choses étant égales par ailleurs, plus on choisit d'éléments par UPE, plus l'effet du plan de sondage est grand (l'effet du plan de sondage est le ratio de la variance d'échantillonnage d'un plan de sondage donné sur celle d'un échantillon aléatoire simple comprenant le même nombre d'éléments). Pour calculer l'effet du plan de sondage, le modèle le plus souvent utilisé est $Deff = 1 + \rho (b - 1)$, où $Deff$ ("Design effect") est l'effet du plan de sondage, ρ , la corrélation intragrappe pour la statistique considérée et b , le nombre d'éléments de l'échantillon par UPE. D'autres études ont montré que, pour beaucoup de variables portant sur l'ensemble de la population aux États-Unis, la corrélation intragrappe observée avec les séries de 100 numéros consécutifs est la plupart du temps inférieure à celle qu'on trouve en général entre les grappes d'échantillons probabilistes aréolaires (voir Groves 1978). Il est possible que cela ne soit pas le cas pour la population noire avec les séries de 100 numéros consécutifs; par ailleurs aucune estimation empirique de la corrélation intragrappe n'a encore été faite avec des séries de 200 numéros consécutifs. L'hypothèse formulée avant d'estimer les erreurs d'échantillonnage était que la corrélation intragrappe devrait avoir la même valeur pour les séries de 100 comme de 200 numéros consécutifs. Cette hypothèse reflète la conception que nous avons formulée plus haut de l'attribution des numéros de téléphone à l'intérieur des circonscriptions téléphoniques.

Suivant les erreurs d'échantillonnage dont la valeur a été estimée à partir d'un ensemble de données de l'étude de production, la valeur moyenne de l'effet du plan de sondage obtenue pour sept statistiques choisies a été de 1.28 avec les séries de 100 numéros consécutifs et de 1.30 avec les séries de 200 numéros consécutifs. La valeur moyenne de l'effet du plan de sondage obtenue pour les séries de 200 numéros consécutifs a été estimée à partir des seuls cas qui tombaient dans la même série de 100 numéros consécutifs que le numéro primaire, tandis que tous les cas ont été utilisés pour calculer la valeur moyenne de l'effet du plan de sondage pour les séries de 200 numéros consécutifs. Ainsi, la taille moyenne des grappes des interviews complètes est de 2.0 dans le cas des séries de 100 numéros consécutifs (coefficient de variation de .043) et de 3.4 pour les séries de 200 numéros consécutifs (coefficient de variation de .029). Ces effets du plan du sondage reflètent la stratification, le regroupement par grappes et la pondération prévues dans le plan de sondage et aussi le fait que la taille des grappes prélevées dans les séries de 100 numéros consécutifs varie plus. (Une des règles de rejet obligeait à prélever le même nombre de ménages noirs dans les séries de 200 numéros consécutifs, mais pas nécessairement le même nombre dans les séries de 100 numéros consécutifs.) Etant donné que les effets du plan de sondage obtenus avec les séries de 100 et les séries de 200 sont très proches l'une de l'autre (1.28 et 1.30), il semble que l'élément

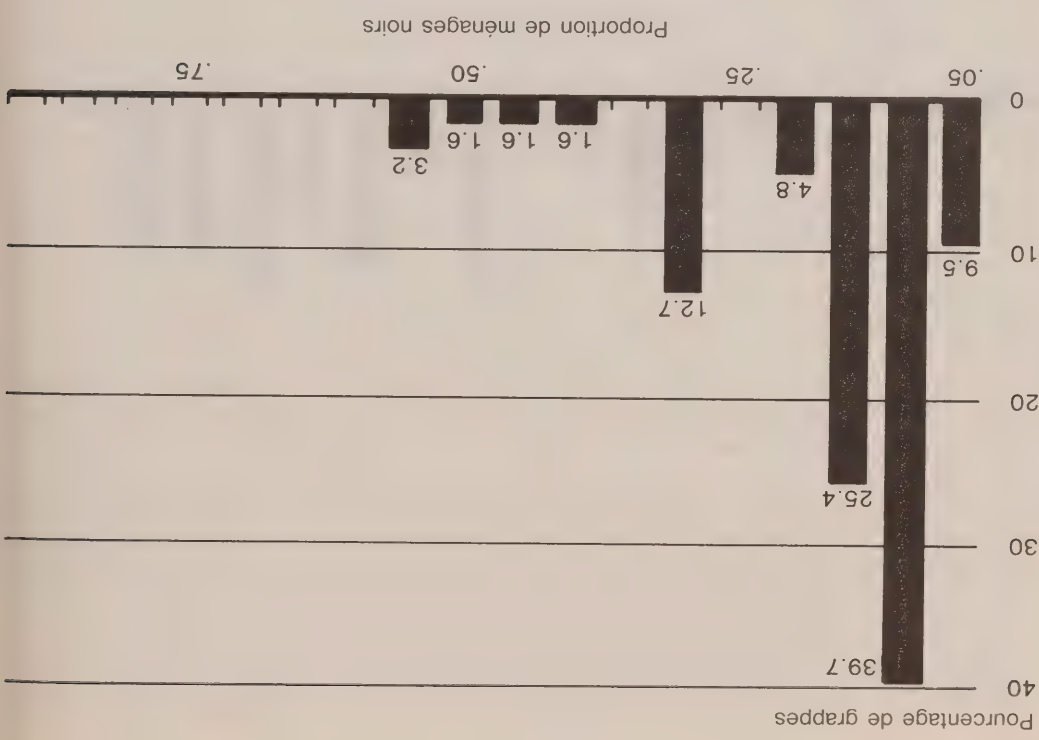


Figure 3. Pourcentage de grappes à faible densité selon la proportion de ménages noirs

La distribution des grappes dans les strates à faible densité et à densité moyenne est très asymétrique, 60 pour cent des UPE de la strate à densité moyenne ayant de 5 à 20 pour cent de ménages noirs. Ces taux d'admissibilité correspondent à un maximum de 10 à 40 ménages noirs pour les UPE de 200 numéros consécutifs prélevées dans la strate à faible densité et la strate à densité moyenne. Dans l'étude de production, la strate à faible densité contenait plusieurs UPE qui n'auraient pas permis de former des grappes de cette taille. (On a estimé que 6 des 63 UPE de cette strate avaient moins de 10 ménages noirs.) La distribution des grappes dans la strate à forte densité est beaucoup plus uniforme. (On a estimé que seulement 4 des 224 UPE de cette strate avaient moins de 10 ménages noirs.)

Ces distributions du pourcentage de ménages noirs par UPE méritent plus d'explications. Compte tenu de ce que nous savons déjà de l'attribution des numéros résidentiels aux banques de numéros disponibles, il n'y a pas lieu de croire qu'à l'intérieur d'une circonscription téléphonique donnée (ou d'un indicatif de central donné), il y ait une tendance générale à attribuer des zones résidentielles différentes à des séries de 100 numéros consécutifs différentes. Cela signifie qu'à l'intérieur d'une circonscription desservant aussi bien des ménages noirs que des ménages non noirs, tout semble indiquer qu'il faut presque nécessairement faire l'hypothèse que les numéros sont attribués sans égard pour la race de l'abonné. Autrement dit, à moins que les circonscriptions ne soient subdivisées en zones rattachées à des noeuds de câbles correspondant à des zones résidentielles de ménages noirs, il n'y a pas de raison *a priori* de supposer que les grappes de ménages noirs dans les séries de 200 numéros consécutifs sont grandes. Suivant ce raisonnement, la distribution plus uniforme des grappes dans la strate à forte densité reflète, croyons-nous, le fait que la proportion de ménages noirs a tendance à varier entre les populations d'abonnés au téléphone des différentes circonscriptions téléphoniques composant la strate.

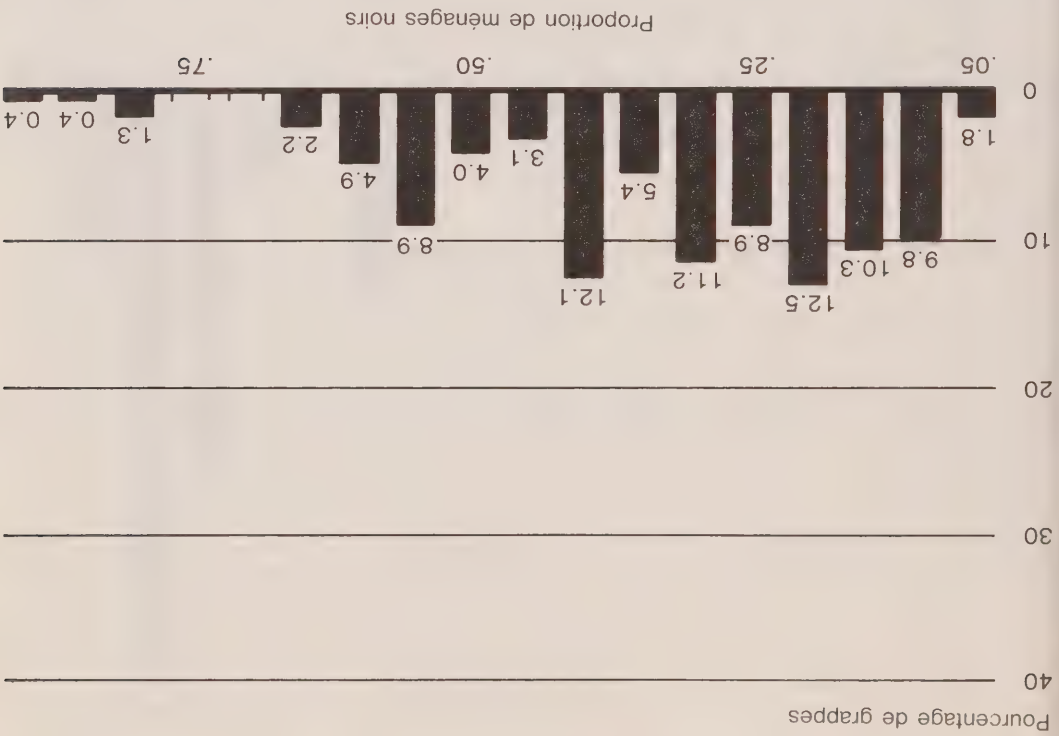


Figure 1. Pourcentage de grappes à forte densité selon la proportion de ménages noirs

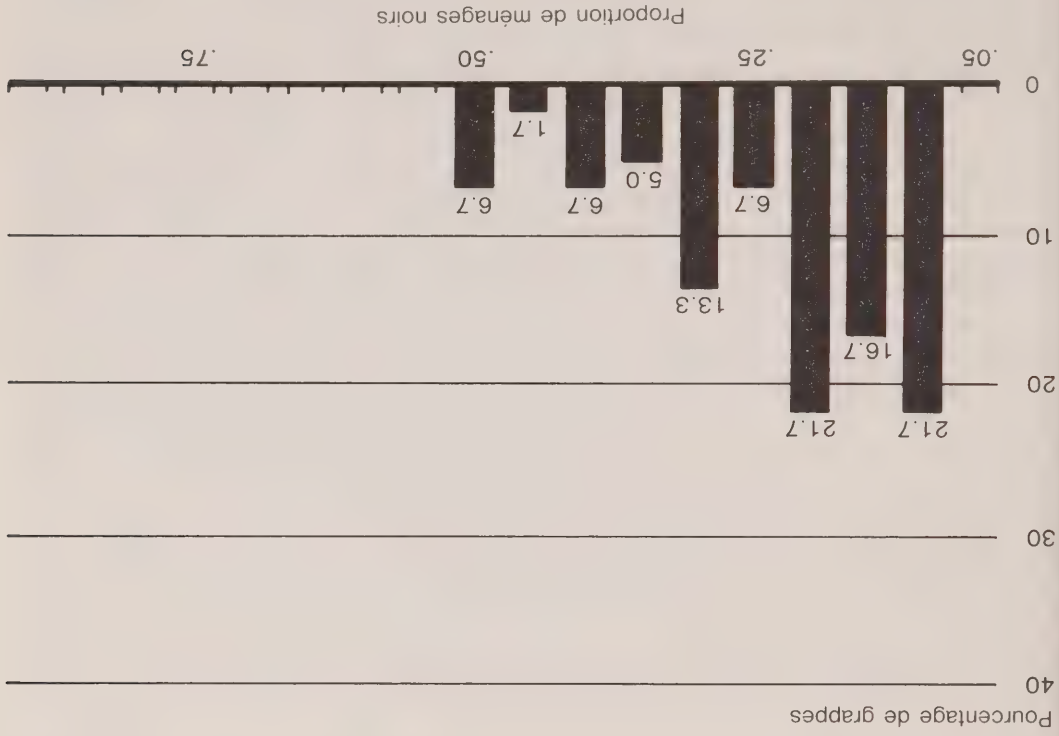


Figure 2. Pourcentage de grappes à densité moyenne selon la proportion de ménages noirs

Comme dans l'étude pilote, le pourcentage de ménages noirs rejoins est différent pour chacune des trois strates, encore que la distinction entre la strate à densité moyenne et la strate à faible densité soit la plus intéressante. Le pourcentage de ménages noirs rejoins à l'aide des numéros secondaires varie approximativement dans des proportions de 2, 1.5 et 1 entre les trois strates. Les trois strates diffèrent également entre elles par la proportion, de numéros de téléphone attribués à des résidences. La strate à forte densité de population noire a proportionnellement plus de numéros secondaires attribués à des unités non résidentielles, ce qui reflète probablement le fait qu'une plus grande partie des circonscriptions téléphoniques de cette strate est urbaine.

Chaque UPE de 200 numéros consécutifs peut être considérée comme deux moitiés d'UPE de 100 numéros consécutifs chacune. Le tableau 3 montre que la proportion de numéros non résidentiels qu'on trouve dans les 100 premiers numéros consécutifs des UPE dans lesquelles le numéro primaire tombait est plus faible que celle qu'on trouve dans l'autre moitié (.378 comparativement à .409), mais cette différence n'est pas statistiquement significative au niveau de .05 (écart-type de .02 environ). De même, la proportion de ménages noirs rejoins à l'aide des numéros secondaires est un peu plus élevée lorsque les numéros secondaires tombaient dans la première moitié de 100 numéros consécutifs du numéro primaire que lorsqu'ils tombaient dans la série de 100 numéros consécutifs adjacente (.133 comparativement à .125). Mais il est peu probable qu'on observe la même différence dans la plupart des cas où cette expérience est répétée. Le tableau 3 donne un autre point de vue sur les résultats du tableau 2; il montre que la réduction observée entre la proportion des ménages admissibles obtenue à partir des séries de 100 numéros consécutifs dans lesquelles le numéro primaire tombe et la proportion obtenue à partir des séries de 100 numéros consécutifs adjacentes est très négligeable.

Le taux d'admissibilité moyen obtenu sur l'ensemble des UPE, qui correspond à la proportion de ménages noirs rejoins, ne devrait pas être le seul critère d'évaluation des plans de sondage. Pour appliquer un plan de sondage "stratifié, chaque UPE retenue doit avoir un nombre suffisant de ménages noirs pour permettre de fixer le nombre de ménages noirs souhaité au second degré. Par conséquent, la distribution de la proportion des numéros admissibles sur l'ensemble des UPE est également importante. Les histogrammes des figures 1, 2 et 3 montrent la distribution de la proportion des ménages noirs par strate sur l'ensemble des UPE. La stabilité des trois distributions varie parce qu'il y a environ quatre fois plus d'UPE dans la strate à forte densité que dans les deux autres (224 UPE dans la strate à forte densité contre 60 environ dans la strate à densité moyenne et la strate à faible densité). Par ailleurs, la forme de la courbe de distribution semble très différente selon la strate.

Tableau 3

Etude de production
Répartition des numéros secondaires selon qu'ils tombent dans la même série de 100 numéros consécutifs que le numéro primaire ou dans la série de 100 numéros consécutifs adjacente

Situation		Répartition	
Même série de 100 numéros consécutifs	Série de 100 numéros consécutifs adjacente	Ménages noirs	Ne connaît pas la race
		.133	.125
Numéros non résidentiels/non valides	Nombre de cas	Ménages non noirs	
		.465	.444
		.378	.409
		(6,522)	(6,511)

secondaires étaient des numéros de ménages noirs (l'écart-type pour cette estimation est de 6 pour cent). Cela est très proche du taux d'admissibilité des numéros secondaires de 12 pour cent enregistré dans l'étude pilote. La comparaison des résultats obtenus au premier degré de sélection avec les résultats obtenus au second degré de sélection montre les gains importants qu'on peut réaliser en utilisant un plan de sondage à deux degrés pour faire des enquêtes téléphoniques auprès des ménages noirs. Les gains à tirer d'un plan de sondage à deux degrés sont le plus marqués dans la strate à faible densité de population noire, où la proportion de ménages rejoints augmente de presque neuf fois entre le premier et le second degré (de .011 à .090). Dans la strate à forte densité, l'augmentation est plus proche du double (.072 à .190). Quant au cas où le plan de sondage utilisé est un plan de sondage à répartition non proportionnelle, les proportions non pondérées de ménages noirs rejoints aux deux degrés de sélection sont de 3 pour cent (au premier degré) et de 15 pour cent (au second degré). Si l'on compare ces chiffres aux estimations obtenues selon le plan de sondage "epsem" (c'est-à-dire à 2 et à 13 pour cent), on constate que les coûts de filtrage sont moins élevés dans le premier cas.

Tableau 2

Etude de production
Répartition des numéros choisis selon la strate

Strate et répartition	Primaires	Secondaires
-----------------------	-----------	-------------

Strate à forte densité de Noirs	Ménages noirs Ne connaît pas la race Ménages non noirs Numéros non résidentiels/non valides	.072 .035 .219 .674	(3,128)	.190 .027 .332 .431	(6,671)
Strate à densité moyenne de Noirs	Ménages noirs Ne connaît pas la race Ménages non noirs Numéros non résidentiels/non valides	.032 .020 .188 .760	(1,879)	.141 .018 .469 .372	(2,375)
Strate à faible densité de Noirs	Ménages noirs Ne connaît pas la race Ménages non noirs Numéros non résidentiels/non valides	.011 .019 .199 .771	(6,116)	.090 .023 .505 .382	(3,987)
Plan de sondage "epsem"*	Ménages noirs Ne connaît pas la race Ménages non noirs Numéros non résidentiels/non valides	.021 .021 .200 .758	(11,123)	.129 .023 .454 .394	(13,033)
Proportion de ménages noirs selon un plan de sondage à répartition non proportionnelle		.031		.150	
Nombre de cas					

* Estimations pondérées des taux obtenus selon un "plan de sondage epsem". Les coefficients de pondération compensent le fait que des fractions de sondage non proportionnelles ont été utilisées pour choisir l'échantillon de l'étude de production à partir des trois strates de densités différentes.

Tableau 1
Etude pilote
Répartition des numéros secondaires choisis dans les séries de
100, 200 et 400 numéros consécutifs par strate

Strate et répartition			
Proportion de tous les numéros choisis			
Strate à forte densité de Noirs			
Séries	Séries	Séries	Séries
de 100	de 200	de 400*	
.205	.201	.214	Ménages noirs
.028	.029	.032	Ne connaît pas la race
.316	.279	.275	Ménages non noirs
.451	.491	.479	Numéros non résidentiels/non valides
(395)	(806)	(1163)	Nombre de cas
Strate à densité moyenne de Noirs			
.104	.080	.076	Ménages noirs
.030	.018	.020	Ne connaît pas la race
.494	.443	.420	Ménages non noirs
.372	.459	.484	Numéros non résidentiels/non valides
(231)	(560)	(878)	Nombre de cas
Strate à faible densité de Noirs			
.085	.084	.069	Ménages noirs
.014	.028	.027	Ne connaît pas la race
.532	.577	.607	Ménages non noirs
.369	.311	.297	Numéros non résidentiels/non valides
(141)	(286)	(491)	Nombre de cas
Total			
.134	.124	.115	Ménages noirs
.024	.025	.026	Ne connaît pas la race
.442	.431	.448	Ménages non noirs
.400	.420	.411	Numéros non résidentiels/non valides
(767)	(1652)	(2532)	Nombre de cas

* Estimation pondérée pour tenir compte du fait que les 9 numéros secondaires des grappes obtenues à partir des séries de 400 numéros consécutifs ne sont pas répartis proportionnellement entre chacune des 4 tranches de 100 numéros consécutifs.

de différence appréciable entre les taux de validité pour les UPE de 200 et de 400 numéros consécutifs, on a opté pour la prudence en choisissant les séries de 200 numéros consécutifs dans l'étude de production. La taille attendue des grappes au second degré a été fixée à 5.5 ménages noirs (sans compter le numéro primaire). Les règles de rejet au premier et au second degré appliquées dans la version modifiée du plan de sondage à deux degrés de Waksberg-Mitofsky ayant servi dans l'étude pilote ont aussi été appliquées dans l'étude de production. Comme des tailles d'échantillon beaucoup plus grandes ont été utilisées dans l'étude de production, on peut répondre aux questions concernant la précision et l'efficacité relative du plan utilisé avec plus d'assurance.

Le tableau 2 présente les résultats du filtrage des numéros primaires et secondaires effectués dans l'étude de production. D'après les taux estimatifs pondérés non biaisés obtenus suivant un plan de sondage "epsem" par CNH à deux degrés, 13 pour cent de tous les numéros

Pour tester la possibilité d'augmenter la taille des UPE, les responsables de l'étude pilote ont prélevé des numéros secondaires dans chacune de ces trois séries. La taille des grappes de ménages noirs du second degré a été fixée à 3 pour les grappes constituées à partir de chaque série de 100 numéros consécutifs associée à un numéro primaire, à 6 pour les grappes constituées à partir de chaque série de 200 numéros consécutifs et à 9 pour les grappes constituées à partir de chaque série de 400 numéros consécutifs. Au premier comme au second degré de sélection, on supposait, si la race du ménage n'était pas connue, qu'il ne s'agissait pas d'un ménage noir.

Le tableau I présente la répartition des numéros secondaires selon le type d'UPE et la strate. Ce qui nous intéresse le plus, c'est la proportion de numéros secondaires attribués aux ménages noirs pour les différentes définitions de l'UPE. Pour les séries de 100 numéros consécutifs, .134 de tous les numéros secondaires sont des numéros rejoignant des ménages noirs. Cela entraîne que .223 de tous les ménages échantillonnés étaient noirs, comparative-ment à la proportion de .25 obtenue par Blair et Czaja. Pour les UPE de 200 numéros consécutifs, .124 de tous les numéros secondaires étaient des numéros rejoignant des ménages noirs, tandis que pour les UPE de 400 numéros consécutifs, .115 de tous les numéros de téléphone prélevés à la deuxième étape étaient attribués à des ménages noirs. L'écart entre ces proportions est toujours inférieur à l'erreur d'échantillonnage (l'écart-type de chaque estimation est d'au moins .02). Cela signifie qu'en portant de 100 à 400 numéros consécutifs la longueur des UPE, on ne diminue pas sensiblement la proportion des numéros rejoignant des ménages noirs (numéros admissibles). Ces taux entraînent qu'étant donné que les UPE de 100 numéros consécutifs permettent d'obtenir en moyenne au second degré des grappes de 13 ou 14 ménages noirs, les séries de 400 numéros consécutifs devraient permettre en moyenne d'obtenir des grappes de 46 ménages noirs. La possibilité d'augmenter la taille des grappes de ménages noirs au second degré d'échantillonnage permet aux chercheurs de réduire considérablement les coûts de filtrage.

Le tableau I permet également de comparer les proportions de numéros secondaires admissibles échantillonnées à partir des trois différentes strates utilisées dans l'étude pilote et selon les trois définitions des UPE retenues. Pour toutes les définitions des UPE retenues (100, 200 et 400 numéros consécutifs), on obtient le même résultat – les circonscriptions téléphoniques des grandes RMSP auxquelles correspond la strate à forte densité de population noire produisent des taux d'admissibilité presque deux fois plus élevés que le taux obtenu pour l'ensemble de la population (.21 contre .12 ou .13). La strate à densité moyenne, correspondant à la strate à faible densité, produit également des taux inférieurs à la moyenne (oscillant entre .07 et .085). Comme la strate à forte densité couvre environ 36 pour cent des ménages noirs abonnées au téléphone, la stratification choisie conjuguée avec la répartition non proportionnelle de l'échantillon au premier degré est un moyen efficace pour réduire les coûts de filtrage.

3. L'ETUDE DE PRODUCTION

L'étude de production a utilisé le plan de stratification élaboré et testé dans l'étude pilote. Un échantillon de 11,223 numéros primaires répartis de façon non proportionnelle a été tiré des trois strates selon la densité de la population noire à l'aide de fractions d'échantillon-nage dans des proportions de 3, 2 et 1 (3 pour la strate à forte densité, 2 pour la strate à densité moyenne et 1 pour la strate à faible densité). Même si l'étude pilote n'a pas révélé

téléphoniques effectuées par le Survey Research Center. Elles donnent à penser que l'extension de la définition des UPE de 100 numéros consécutifs à un nombre plus élevé pourrait nous permettre d'utiliser des grappes de numéros secondaires plus grandes sans beaucoup réduire la proportion de ces numéros qui rejoignent des ménages noirs.

2. L'ÉTUDE PILOTE

Dans deux expériences intégrées effectuées dans le cadre d'une enquête pilote, plusieurs plans de sondage différents ont été testés. Un des objectifs de l'étude pilote était de déterminer dans quelle mesure on peut produire des séries de numéros de téléphone à forte densité de numéros d'abonnés noirs au moyen d'une stratification fondée sur les régions administratives et d'une correspondance approximative entre les limites des régions administratives et celles des circonscriptions téléphoniques. Pour cela, trois strates de circonscriptions ont été définies:

- 1) "À forte densité" – Circonscriptions téléphoniques correspondant aux villes centrales des grandes régions métropolitaines statistiques normalisées (par exemple la ville de Chicago pour la RMSN de Chicago). Cette définition était fondée sur le nom des circonscriptions de ces régions.
- 2) "À densité moyenne" – Toutes les autres circonscriptions dans certains Etats choisis du sud des Etats-Unis (Virginie, Caroline du Nord, Caroline du Sud, Floride, Georgie, Alabama, Mississippi et Louisiane). La très grande majorité des circonscriptions de ces Etats sont comprises dans un seul Etat; celles qui desservaient deux Etats en même temps ont été associées à l'Etat dont le nom apparaît dans le nom de la circonscription.
- 3) "À faible densité" – Toutes les autres circonscriptions des Etats limitrophes des Etats-Unis.

Un échantillon avec équiprobabilité de sélection de 1,400 combinaisons à six chiffres dont trois pour l'indicateur régional et trois pour l'indicateur du central a ensuite été prélevé systématiquement à partir des 34,389 combinaisons de ce genre qui sont inscrites comme étant en service dans une base qu'on peut acheter de la société American Telephone & Telegraph (AT&T). Des numéros aléatoires à 4 chiffres ont été ajoutés à chacun des radicaux à six chiffres choisis pour produire un échantillon de 1,400 numéros primaires à dix chiffres. Les résultats de l'étude pilote ont montré que la proportion de numéros de téléphone rejoignant des ménages noirs varierait beaucoup selon la strate. On a constaté qu'il fallait faire six fois plus de filtrage pour trouver un ménage noir dans la strate à faible densité que dans la strate à forte densité. (Ce résultat a été confirmé avec plus de précision dans l'étude de production, dont nous parlerons dans la prochaine section).

Un autre objectif de l'étude pilote était de tester l'utilisation de règles de rejet fondées sur la composition raciale des ménages correspondant aux numéros de téléphone de l'échantillon tiré des UPE de différentes tailles et sur la validité/non validité de ces numéros (suivant qu'ils rejoignent ou non un ménage). Pour augmenter la précision des analyses relatives à cet objectif, 500 autres numéros primaires ont été choisis dans les strates à forte et à moyenne densité. Les 1,900 numéros primaires de l'échantillon combiné de l'étude pilote ainsi prélevé ont ensuite été composés pour repérer les ménages noirs. Quand un numéro primaire échantillonné rejoignait un ménage noir, il correspondait en même temps à trois UPE différentes. Comme on peut le voir au tableau 1, chaque numéro individuel peut être considéré comme faisant partie en même temps d'une série de 100 numéros consécutifs, d'une série de 200 numéros consécutifs et d'une série de 400 numéros consécutifs. Par exemple, le numéro 313-764-4424 fait partie de la série des 100 numéros consécutifs 4400-4499, de la série des 200 numéros consécutifs 4400-4599 et de la série des 400 numéros consécutifs 4400-4799.

La stratification de la population des ménages abonnés au téléphone selon la race a pour objet d'isoler les circonscriptions (zones de rattachement à un central) à forte proportion d'abonnés de race noire. On applique ensuite à ces strates des fractions d'échantillonnage plus grandes que celles qui sont appliquées aux strates qui ont une plus faible proportion de ménages noirs. Selon ce plan de sondage à répartition non proportionnelle, le nombre total de ménages qu'il faut rejoindre pour obtenir une interview avec un ménage noir admissible est plus petit que le nombre de ménages qu'il faudrait rejoindre si l'on utilisait un échantillon "epsem" de la population des ménages. Les coûts de filtrage qu'entraîne la constitution d'un échantillon de ménages noirs sont ainsi diminués. Dans les échantillons d'enquêtes téléphoniques, l'unité géographique de base est la circonscription correspondant à un central téléphonique et à laquelle un ou plus d'un indicatif à trois chiffres est attribué. En général, le compte des abonnés n'est pas ventilé selon les caractéristiques raciales. Il faut par conséquent utiliser des indicateurs approximatifs des circonscriptions à forte densité de population noire. Les expériences décrites dans le présent document avaient pour objet notamment d'examiner la valeur de ces indicateurs d'approximation.

Blair et Czaja (1982) ont élaboré une variante du plan de sondage par CNH de Waksberg-Mitofsky qui intègre des règles de rejet à deux degrés fondées aussi bien sur le fait que le numéro composé est un numéro résidentiel ou non et sur la race du ménage. Appliquée à la population noire, cette méthode inclut à la première étape seulement des séries de 100 numéros consécutifs dont le numéro primaire a été attribué à un ménage noir et prélève ensuite dans chacune des UPE un nombre fixe d'avance de numéros de téléphone de ménages noirs. Dans une enquête nationale par sondage aux Etats-Unis, Blair et Czaja ont constaté qu'en utilisant ce plan, le pourcentage de ménages noirs rejoints à l'aide de tous les numéros de téléphone choisis augmentait à 25 pour cent à la deuxième étape, tandis qu'il était de 9 pour cent à la première étape. Etant donné les probabilités de sélection qui se compensent dans les deux étapes, ce plan de sondage "epsem" réduit considérablement le travail de filtrage requis pour obtenir un échantillon de taille donnée de ménages noirs. Une variante analogue des règles de rejet utilisées dans le plan de sondage à deux degrés de Waksberg-Mitofsky a été utilisée dans les expériences décrites dans le présent document.

Dans le plan de sondage de Blair et Czaja, certaines séries de 100 numéros consécutifs de la première étape ne contenaient pas assez de numéros de téléphone de ménages noirs pour produire le nombre d'éléments par grappe (10 dans ce cas-là) qu'il faut pour constituer un échantillon "epsem" de ménages noirs. En outre, ce plan comporte des coûts assez élevés de filtrage au premier degré de sélection; il faut composer plus de 44 numéros primaires pour rejoindre un ménage noir. La solution commune à ces deux problèmes est d'augmenter la taille de l'UPE et de choisir plus d'éléments dans les UPE à la deuxième étape. Les analyses présentées ici ont utilisé des unités de sondage du premier degré de 100, 200 et 400 numéros consécutifs chacune. L'extension de la définition des UPE pour qu'elles comprennent plus de numéros que la norme habituelle de 100 numéros consécutifs est suggérée par les observations sur la répartition des numéros de téléphone à l'intérieur d'un indicatif donné. Voici les principales constantes qui semblent se dégager de nos observations: 1) presque tous les numéros de ménages dont le numéro de téléphone commence par un indicatif donné desservent des unités situées à l'intérieur des limites géographiques de la circonscription téléphonique correspondante à l'indicateur en question; 2) il y a peu de regroupements géographiques par grappes des numéros attribués à l'intérieur des limites géographiques des circonscriptions téléphoniques (c'est-à-dire qu'en général deux ménages voisins n'ont pas deux numéros consécutifs et n'ont pas nécessairement deux numéros comportant le même indicatif); et 3) le pourcentage des numéros rejoignant des ménages varie plus d'une série de 1,000 numéros consécutifs à une autre qu'entre les séries de 100 numéros consécutifs d'une même série de 1,000 numéros consécutifs. Ces constatations sont le fruit de plusieurs années d'enquêtes

L'utilisation de techniques d'échantillonnage et d'interview par téléphone exclut de l'enquête les Noirs faisant partie de ménages qui n'ont pas le téléphone (soit 15% environ de toute la population noire aux États-Unis). Dans la plupart des cas, les personnes non abonnées au téléphone sont plus pauvres et plus jeunes que celles qui vivent dans les ménages abonnées au téléphone (Thornderry et Massey 1983). Dans la mesure où il est possible que les Noirs non abonnées au téléphone n'aient ni les mêmes attitudes politiques ni les mêmes intentions de vote que ceux qui sont abonnés, les estimations produites à partir des résultats de l'enquête risquent de différer des vrais paramètres de l'ensemble de la population noire des États-Unis. Sans chercher à minimiser l'erreur de non couverture associée aux enquêtes téléphoniques auprès de la population noire, nous sommes principalement attachés aux différences au niveau de l'efficacité sur le plan des coûts et de l'erreur d'échantillonnage pouvant résulter de l'utilisation d'approches alternatives dans la formation des échantillons de ménages noirs pour des enquêtes téléphoniques.

Les plans de sondage pour enquêtes téléphoniques présentés ici ont été élaborés à partir d'un plan défini par Waksberg (1978). Ce plan de sondage par composition de numéros de téléphone au hasard (CNH), communément appelé méthode de Waksberg-Mitofsky, est un plan d'échantillonnage de numéros de téléphone par grappes à deux degrés. Les numéros de téléphone aux États-Unis comprennent 10 chiffres, soit un code régional à 3 chiffres, un code ou indicatif de central à trois chiffres et un numéro de ligne à 4 chiffres de 0000 à 9999, par exemple 313-764-4424. À la première étape, un échantillon stratifié de numéros de téléphone à dix chiffres est produit au hasard et chaque "numéro primaire" ainsi produit est lié à un bloc de 100 numéros consécutifs (par exemple le numéro 313-764-4424 serait lié à la série de 100 numéros consécutifs 313-764-4400 à 313-764-4499). Si, dans le cas des enquêtes-ménages, il se trouve que le numéro primaire est un numéro de téléphone de ménage valide, alors la série de 100 numéros de téléphone consécutifs dont il fait partie est retenue à la première étape pour une deuxième étape d'échantillonnage. Sinon, sa série de 100 numéros consécutifs est rejetée. Par conséquent, la probabilité de sélection d'une série de 100 numéros consécutifs à la première étape est proportionnelle au nombre de numéros de téléphone de ménage valides dans cette série. À la deuxième étape de l'échantillonnage, des numéros consécutifs retenues à la première étape. Par conséquent, l'échantillonnage des numéros consécutifs, à la deuxième étape, est un échantillonnage avec probabilité conditionnelle de sélection inversement proportionnelle au nombre de numéros de ménage valides dans chaque série de 100 numéros consécutifs. Le plan produit ainsi un échantillon de numéros de téléphone de ménage avec équiprobabilité de sélection ("epsem sampling") et regroupe ces numéros par grappes de sorte que la proportion du nombre total de numéros de téléphone sélectionnés qui rejoignent un ménage est plus grande que celle qui aurait été obtenue au moyen d'un échantillonnage stratifié par CNH. Par souci de clarté dans l'exposé qui va suivre, nous appellerons les banques de 100 numéros de téléphone consécutifs, unités de sondage du premier degré ou unités primaires d'échantillonnage (UPE) du plan de sondage par CNH à deux degrés. Le terme "grappe" est réservé pour désigner l'ensemble comprenant un nombre fixé de numéros de téléphone de ménage valides choisis à partir des UPE à la deuxième étape du plan de sondage.

Dans la présente étude, les modifications apportées au plan de sondage pour réduire les coûts de filtrage prennent trois formes: a) stratification des circonscriptions téléphoniques selon la proportion de Noirs et prélèvement non proportionnel (c'est-à-dire d'une proportion plus élevée) d'une partie de l'échantillon dans les strates à forte densité de population noire; b) utilisation de règles de rejet à deux degrés fondées sur le fait que le numéro de téléphone composé est un numéro résidentiel ou non et sur la race du ménage; et c) augmentation de la taille des UPE (de 100 numéros consécutifs à 200 et 400).

Plans de sondage d'enquêtes téléphoniques auprès de ménages noirs aux États-Unis¹

KATHRYN M. INGLIS, ROBERTS M. GROVES, et STEVEN G. HEERINGA²

RÉSUMÉ

Le plan de sondage d'enquêtes téléphoniques à deux degrés avec règles de rejet élaboré par Wakseberg en 1978 a été modifié pour améliorer l'efficacité des enquêtes téléphoniques auprès de la population noire aux États-Unis. L'application expérimentale d'autres plans de sondage a montré que: a) l'utilisation d'une stratification approximative fondée sur les noms des circonscriptions téléphoniques et des États; b) l'utilisation au premier degré de définitions de séries de 200 et 400 numéros consécutifs déterminant des grappes de grande taille; et c) l'application de règles de rejet fondées sur la race des ménages sont trois éléments dont la réunion améliore la précision relative d'un échantillon, étant donné des ressources fixes. Des modèles de coûts et des modèles d'erreur ont aussi été examinés pour simuler l'application d'autres plans de sondage.

MOTS CLÉS: Échantillons par CNH; enquêtes téléphoniques; échantillons de population rare.

1. INTRODUCTION

Les enquêtes effectuées auprès de populations rares et qui n'ont pas de base spéciale coûtent souvent plus cher par unité que celles qui visent une population complète. Quand la population rare est un petit sous-groupe d'une population facile à définir, l'échantillon de ce sous-groupe est souvent obtenu par filtrage de la population plus grande. Les enquêtes auprès de ménages de sous-groupes démographiques comme la population noire des États-Unis, procèdent en général de cette façon pour trouver les unités de sondage susceptibles d'être incluses dans l'échantillon; toutefois l'utilisation à grande échelle de techniques de filtrage pour constituer des échantillons de population rare coûte cher par interview. Ces dernières années, des méthodes de sondage par téléphone ont été proposées comme moyens peu coûteux pour échantillonner et interviewer des populations rares. Le coût d'une interview téléphonique est souvent inférieur à celui d'une interview en personne (Groves et Kahn 1979) et lorsqu'il faut absolument appliquer une méthode de filtrage pour déterminer les répondants susceptibles d'entrer dans un échantillon, l'efficacité des interviews téléphoniques en termes de coûts est encore plus marquée. Tout de même, les coûts de filtrage dans les enquêtes téléphoniques auprès de populations rares peuvent être élevés en chiffres absolus. Dans le présent document, nous présentons diverses façons d'affiner la méthode de filtrage dans les enquêtes téléphoniques pour réduire les coûts tout en obtenant le degré de précision voulu. Pour cela, nous avons examiné divers plans de sondage qui pourraient s'appliquer à des enquêtes téléphoniques auprès de la population noire des États-Unis. Les essais d'enquête téléphonique décrits ici ont été effectués dans le cadre d'un sondage d'opinion mené auprès des Noirs au moment de la campagne électorale de 1984 pour la présidence des États-Unis afin de connaître leurs attitudes politiques et leurs intentions de vote.

¹ Une version révisée d'un article qui a été présenté au congrès de 1985 de l'American Statistical Association. La recherche a été partiellement subventionnée par l'U.S. Bureau of the Census et le Survey Research Center. La discussion ne représente pas nécessairement les opinions de ces organismes-là.

² Kathryn M. Inglis, McNair Anderson and Associates, Australie; Robert M. Groves et Steven G. Heeringa, Survey Research Center, University of Michigan, Ann Arbor, Michigan, 48106-1248, États-Unis.

TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada
Volume 13, numéro 1, juin 1987

TABLE DES MATIÈRES

K.M. INGLIS, R.M. GROVES, et S.G. HEERINGA	Plans de sondage d'enquêtes téléphoniques auprès de ménages noirs aux États-Unis.....	1
W.M. SYKES et M. COLLINS	Comparaison entre l'interview téléphonique et l'interview sur place au Royaume-Uni.....	19
G.J. BRACKSTONE	Utilisation des dossiers administratifs à des fins statistiques	35
C.A. FRANCISCO, W.A. FULLER, et R. FECESO	Propriétés statistiques des estimateurs de la production végétale	53
E.B. DAGUM	Problèmes courants sur la désaisonnalisation	71
S. KUMAR et A.C. SINGH	Sur l'estimation efficace des taux de chômage à l'aide de données de l'enquête sur la population active	83
A. DEY et A.K. SRIVASTAVA	Méthode d'échantillonnage avec probabilités de sélection proportionnelles à la taille	93
D. DOLSON, K. MCLEAN, J.-P. MORIN, et A. THÉBERGE	Plan d'échantillonnage pour l'enquête sur la santé et les limitations d'activités ..	101
F.C. OKAFOR	Comparaison d'estimateurs de totaux de population obtenus par sondages successifs à deux degrés à l'aide de l'information auxiliaire.....	119
	Rectification	133

TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada

La revue Techniques d'enquête est répertoriée dans The Survey Statistician et Statistical Theory and Methods Abstracts. On peut en trouver les références dans Current Index to Statistics

COMITÉ DE DIRECTION

Président

G.J. Brackstone

Membres

B.N. Chinappa

G.J.C. Hole

C. Patrick

F. Mayda (Directeur de la production)

M.P. Singh

D. Roy

R. Platek

COMITÉ DE RÉDACTION

Rédacteur en chef

M.P. Singh, *Statistique Canada*

Rédacteurs associés

K.G. Basavarajappa, *Statistique Canada*

D.R. Bellhouse, *University of Western*

Ontario

L. Biggert, *Université de Florence*

D. Binder, *Statistique Canada*

E.B. Dagum, *Statistique Canada*

W.A. Fuller, *Iowa State University*

J.F. Gentleman, *Statistique Canada*

D. Holt, *University of Southampton*

Rédacteurs adjoints

J. Armstrong, *Statistique Canada*

H. Lee, *Statistique Canada*

POLITIQUE DE RÉDACTION

La revue Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception de l'ordre pratique, l'utilisation de différentes sources de données et de méthode de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

La revue Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à en faire parvenir le texte au rédacteur en chef, M. M.P. Singh, Division des méthodes d'enquêtes sociale, Statistique Canada, 4^e étage, Édifice Jean-Talon, Tunney's Pasture, Ottawa (Ontario), Canada KIA 0T6. Prière d'envoyer deux exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

Abonnement

Le prix de la revue Techniques d'enquête (catalogue n° 12-001) est de 20,00\$ par année au Canada et de 23,00\$ par année à l'étranger (paiement en dollars canadiens ou l'équivalent). Prière de faire parvenir votre demande d'abonnement à: Section des ventes des publications, Statistique Canada, Ottawa (Ontario), Canada KIA 0T6. Un prix réduit, soit 10,00\$ (E.-U.) (\$14,00 Can.) est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête et la Société Statistique du Canada. Veuillez envoyer votre demande d'abonnement directement à l'organisation.

TECHNIQUES D'ENQUÊTE

UNE REVUE DE STATISTIQUE CANADA
JUN 1987

Publication autorisée par
le ministre des Approvisionnements
et Services Canada
©Ministre des Approvisionnements
et Services Canada 1987

Le lecteur peut reproduire sans autorisation des
extraits de cette publication à des fins d'utilisation
personnelle à condition d'indiquer la source en
entier. Toutefois, la reproduction de cette publication
en tout ou en partie à des fins commerciales ou de
redistribution nécessite l'obtention au préalable
d'une autorisation écrite des Services d'édition,
Agent de droit d'auteur, Centre d'édition du gouvernement
du Canada, Ottawa, Canada K1A 0S9.

Novembre 1987

Prix: Canada, \$20.00 par année
Autres pays, \$23.00 par année

Paieement en dollars canadiens ou l'équivalent
Catalogue 12-001, vol. 13, n° 1

ISSN 0714-0045

Ottawa

Canada

VOLUME 13, NUMÉRO 1
JUN 1987

UNE REVUE
DE
STATISTIQUE CANADA

TECHNIQUES D'ENQUÊTE

Statistique Canada Statistics Canada



12
- 001



Statistics Canada Statistique Canada

SURVEY METHODOLOGY

A JOURNAL
OF
STATISTICS CANADA

VOLUME 13, NUMBER 2
DECEMBER 1987

Canada

SURVEY METHODOLOGY

A JOURNAL OF STATISTICS CANADA

DECEMBER 1987

Published under the authority of
the Minister of Supply and
Services Canada

©Minister of Supply
and Services Canada 1988

Extracts from this publication may be reproduced
for individual use without permission provided the
source is fully acknowledged. However, reproduction
of this publication in whole or in part for purposes
of resale or redistribution requires written permission
from the Publishing Services Group, Permissions
Officer, Canadian Government Publishing Centre,
Ottawa, Canada K1A 0S9

May 1988

Price: Canada, \$20.00 a year
Other Countries, \$23.00 a year

Payment to be made in Canadian funds or equivalent

Catalogue 12-001, Vol. 13, No. 2

ISSN 0714-0045

Ottawa

SURVEY METHODOLOGY

A Journal of Statistics Canada

The Survey Methodology Journal is abstracted in The Survey Statistician and Statistical Theory and Methods Abstracts and is referenced in the Current Index to Statistics.

MANAGEMENT BOARD

Chairman G.J. Brackstone

Members B.N. Chinnappa

G.J.C. Hole

C. Patrick

F. Mayda (Production Manager)

R. Platek

D. Roy

M.P. Singh

EDITORIAL BOARD

Editor M.P. Singh, *Statistics Canada*

Associate Editors

K.G. Basavarajappa, *Statistics Canada*

D.R. Bellhouse, *University of Western Ontario*

L. Biggeri, *University of Florence*

D. Binder, *Statistics Canada*

E.B. Dagum, *Statistics Canada*

W.A. Fuller, *Iowa State University*

J.F. Gentleman, *Statistics Canada*

D. Holt, *University of Southampton*

G. Kalton, *University of Michigan*

W.M. Podehl, *Statistics Canada*

M.N. Murthy, *Applied Statistics Centre, India*

J.N.K. Rao, *Carleton University*

I. Sande, *Statistics Canada*

C.E. Särndal, *University of Montreal*

F.J. Scheuren, *U.S. Internal Revenue Service*

V. Tremblay, *Statplus, Montreal*

K.M. Wolter, *U.S. Bureau of the Census*

Assistant Editors

J. Armstrong, J. Gambino and H. Lee, *Statistics Canada*

EDITORIAL POLICY

The Survey Methodology Journal publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

The Survey Methodology Journal is published twice a year. Authors are invited to submit their manuscripts in either of the two official languages, English or French to the Editor, Dr. M.P. Singh, Social Survey Methods Division, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Two nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of the Survey Methodology Journal (Catalogue No. 12-001) is \$20.00 per year in Canada, \$23.00 per year for other countries (payment to be made in Canadian funds or equivalent). Subscription order should be sent to: Publication Sales, Statistics Canada, Ottawa Ontario, Canada K1A 0T6. A reduced price of US \$10.00 (\$14.00 Can.) is available to members of the American Statistical Association, the International Association of Survey Statisticians and the Statistical Society of Canada. Please subscribe through your organization.

SURVEY METHODOLOGY

A Journal of Statistics Canada
Volume 13, Number 2, December 1987

CONTENTS

In this Issue	125
A. GIOMMI	
Nonparametric Methods for Estimating Individual Response Probabilities.....	127
C.S. WITHERS	
Estimates Based on Randomly Rounded Data	135
G.H. CHOUDHRY and H. LEE	
Variance Estimation for the Canadian Labour Force Survey	147
P. ANTOINE, X. BRY, and P.D. DIOUF	
The "AGEVEN" Record: A Tool for the Collection of Retrospective Data	163
Special Section - Estimation and Weighting Methods	
K.R. COPELAND, F.K. PEITZMEIER, and C.E. HOY	
An Alternative Method of Controlling Current Population Survey Estimates to Population Counts.....	173
C.H. ALEXANDER	
A Class of Methods for Using Person Controls in Household Weighting	183
G. LEMAÎTRE and J. DUFOUR	
An Integrated Method for Weighting Persons and Families	199
H.L. OH and F. SCHEUREN	
Modified Raking Ratio Estimation	209
Short Communications	
S.G. PRABHU-AJGAONKAR	
Comparison of the Horvitz-Thompson Strategy with the Hansen-Hurwitz Strategy.....	221
Acknowledgments	225

In This Issue

Two new features appear for the first time in this issue of Survey Methodology. "**In This Issue**" summarizes papers appearing in the Journal and will appear regularly. The other new feature, a "**Short Communications**" section, will appear in the Journal from time to time.

This issue contains nine papers, four dealing with **estimation and weighting methods**, including two on family estimation. Fritz Scheuren's initiative and editorial assistance were instrumental in putting this special section together.

The first three papers in the special section deal (at least in part) with least-squares methods for weighting survey data. There is a certain historical irony in this. In their 1940 paper, Deming and Stephan introduced iterative proportional fitting as a quick practical way for approximating the estimates obtained by minimizing a squared function of the cells of a contingency table, subject to restrictions on the margins. The use of this technique has become fairly generalized in weighting survey data, where it is known as "raking ratio estimation".

In "An Alternative Method of Controlling Current Population Survey Estimates to Population Counts", Copeland, Peitzmeier and Hoy compare a raking ratio estimator to a generalized least-squares estimator under the same marginal restrictions. The comparison is carried out for estimates of individual characteristics obtained from the Current Population Survey, a household survey conducted by the United States Bureau of the Census. They note that the estimates produced by the two methods are very similar.

Most current methods of weighting data from household surveys produce weights that differ from person to person within the same household. A single weight per household, in addition to its conceptual appeal, would eliminate the recurrent and often awkward discrepancies between person-based and family-based estimates. Alexander, in "A Class of Methods for Using Person Controls in Household Weighting", considers a class of "constrained minimum distance" methods (including GLS) which actually yield a single weight per household yet respect person-level marginal totals. The properties of these methods in the presence of undercoverage are then studied through some simple coverage models.

Lemaître and Dufour, in "An Integrated Method for Weighting Persons and Families", propose a regression estimator that also yields a single weight per household and is equivalent to the GLS estimator under certain general conditions. Using Canadian Labour Force Survey data, they obtain large efficiency gains for estimates of families, and marginal gains for estimates of persons, relative to current methods.

In the last paper in this section, "Modified Raking Ratio Estimation", Oh and Scheuren describe an estimation procedure similar to the usual raking ratio. Their method can be used when population totals are available not only for the margins, but also for interior cells in a multi-way table. It combines conventional ratio estimation for cells with large sample sizes and raking ratio estimation for cells with sample sizes that are small (or zero). In an application involving sampling of corporate income tax returns, the Oh-Scheuren approach produced more efficient estimates relative to conventional ratio estimation. The authors stress that, before their method is offered for wide use, further work is needed including, among other things, comparison with conventional collapsing schemes.

The other four papers in this issue consider the development and application of methods and procedures with regard to probabilities of response in a survey context, rounding criteria for protection of confidentiality, data collection and analysis for retrospective type surveys, and variance estimation for the Canadian Labour Force Survey.

Every survey has some nonresponse problems. These are usually handled by imputation or adjustment procedures based on the assumption that nonresponse occurs at random within imputation or adjustment classes. The resulting estimates are generally biased whenever this assumption is not satisfied. Various methods of estimating response probabilities involving models have been proposed, notably by Cassel, Särndal and Wretman (CSW), but these methods are not effective when the assumed model is inadequate. In "Nonparametric Methods for Estimating Individual Response Probabilities", Giommi describes nonparametric procedures for estimating response probabilities using auxiliary information, providing an alternative to the CSW estimator that is robust against both population and response model breakdown. The resulting estimators perform well in Monte Carlo simulation studies.

Random rounding is used to ensure the confidentiality of information about individual in statistical aggregates. In the context of the 1971 Canadian Census, Nargundkar and Saveland developed a rounding process that is unbiased in the sense that the expected value of the rounded data is the same as that of the unrounded data. Fellegi (SMJ, 1975) introduced controlled random rounding, a procedure that, in addition to being unbiased, also preserves additivity. Several other papers have since appeared, including the very recent work of Cox (JASA, 1987), generalizing and extending the applications to other fields. In "Estimates Based on Randomly Rounded Data", Withers develops an expression for the variance of unbiased estimates of cell probabilities and presents a comparison of efficiencies involving the rounding processes used in Australia, the United Kingdom, New Zealand and Canada. He also extends his results to any smooth function of the cell probabilities for applications in different areas of statistics.

In "Variance Estimation for the Canadian Labour Force Survey", Choudhry and Le describe studies conducted to select a variance estimator for raking ratio estimates from the Canadian Labour Force Survey. Their paper reports on a comparison of three variance estimators for the random group sampling design: Keyfitz, Rao-Hartley-Cochran and Rao. In spite of its slight inferiority to the other two methods in terms of bias and stability, the Keyfitz method is suggested for actual use because of its operational simplicity.

In "The 'AGEVEN' Record: A Tool for the Collection of Retrospective Data", Antoine Bry and Diouf describe techniques used to collect data on natality and mortality of women in Pikine, a suburb of Dakar, Senegal. The retrospective procedure employed involved placing observed events (mainly births and deaths) in their socio-economic context and, according to the authors, made it possible to "better assess the relationship between urban insertion and changes in demographic behaviour". Analysis of data from the survey clearly indicates that child mortality rates are higher for children born in rural villages than for those born in Pikine.

It is well known that the Hansen-Hurwitz strategy is inferior to the Horvitz-Thompson strategy associated with a number of IPPS (inclusion probability proportional to size) sampling procedures. In the final piece in this issue, in the "Short Communications" section, Prabh Ajgaonkar presents proofs of these results that are much simpler than those already available in the literature.

Nonparametric Methods for Estimating Individual Response Probabilities

ANDREA GIOMMI¹

ABSTRACT

This paper deals with the nonresponse problem in the estimation of the mean of a finite population, following an approach closely related to that of Cassel, Särndal and Wretman (1983). Two very simple methods are proposed for estimating the individual response probabilities; these are then used, in connection with a superpopulation model, to construct estimators for the population mean. A first evaluation of the properties of the proposed methods is given by a Monte Carlo experiment. The results shed some light on their effectiveness.

KEY WORDS: Nonresponse; Individual response probability; Nonparametric methods.

1. INTRODUCTION

Dealing with the estimation of finite population mean (or total, etc.) in the presence of nonresponse, Cassel, Särndal and Wretman (1983) introduced a very general estimation method based on the fundamental concept of individual response probability (IRP). The authors proposed estimators which are in part determined by a superpopulation model and in part by a response model, i.e., a model formalizing the response mechanism and by which IRP can be estimated from sample data. The estimation of IRP is the crucial point of their theory. In fact, if the superpopulation model is not correctly chosen, as is often the case, only a correct choice of the response model may guard the estimators from design bias. By a Monte Carlo experiment, Giommi (1985a) showed that a response model supplying a "good approximation" of the "true" response model can restore virtual unbiasedness; but little is known about the extent of a good approximation and in any case the choice of a response model may prove cumbersome besides being arbitrary. A natural way of avoiding these difficulties is to estimate the IRP by nonparametric procedures. In the present paper we propose two very simple methods to estimate IRP when available auxiliary information (which is assumed to be related to the response behaviour) is represented by a single continuous variable. The methods which make use of some tools of the kernel estimation theory may be viewed as an extension of the popular correction technique for nonresponse consisting in reweighting units by adjustment cells.

In this paper some empirical evaluations of these methods are described and the results regarding the bias and efficiency of the related estimators are presented.

2. ESTIMATION OF THE INDIVIDUAL RESPONSE PROBABILITIES

Let us consider a population of N units labelled k ($k = 1, 2, \dots, N$), and let Y be a variable under study, of which we want to estimate the mean $\bar{Y} = \sum_k y_k / N$ from a sample s of n units, the selection being based on a given design $p(s)$. For the estimation, auxiliary information is available, represented by known values x_k ($k = 1, \dots, N$), of a scalar continuous

¹ Andrea Giommi, Department of Statistics, University of Florence, Via Curtatone, 1, 50123 Florence, Italy.

variable X (the extension of the procedures proposed for the multidimensional case is, in principle, straightforward).

In the sample, Y is observable only in a subset r of n_r respondents and not on the $n - n_r$ nonrespondents. After the selection of the sample, the available information can be represented as follows:

$$(k, I_k, I_k y_k, x_k) \quad k \in s; N, n,$$

where I_k is an indicator random variable such that $E(I_k) = q_k$ and q_k is the IRP.

To estimate q_k , a parametric model is generally assumed (Cassel *et al.* 1983) such that:

$$q_k = q(\Theta, x_k),$$

where Θ is an unknown parameter (or vector of parameters) and $q(\cdot, \cdot)$ is a functional form to be specified. Estimated q_k are then obtained replacing in the above parametric model estimated values $\hat{\Theta}$ of Θ .

In this paper the estimates of q_k ($k \in r$) are obtained by avoiding any parametric specification of the function $q(\cdot, \cdot)$; nevertheless, maintaining the hypothesis that the IRPs depend on the values x_k . Two procedures (methods (1) and (2)) are proposed.

In the first, q_k ($k \in r$) is estimated as the response rate (i.e. the proportion of respondents) in a group of units centered on the unit k , corresponding to an appropriate interval of x -values centered at x_k . Assuming that $2h_k$ is the length of such an interval, q_k is estimated by the following ratio:

$$\hat{q}_k = \sum_{j \in r} D(x_k - x_j) / \sum_{j \in s} D(x_k - x_j), \quad (1)$$

where

$$D(x_k - x_j) = \begin{cases} 1 & \text{if } |x_k - x_j| \leq h_k \\ 0 & \text{otherwise.} \end{cases}$$

It is evident that the estimate \hat{q}_k depends on h_k or h if we adopt – as in this paper – a constant interval; the numerical specification of h is a main problem in applications.

In the second procedure, all the sample units, rather than a group, contribute to the estimation of q_k . By this method the possible limitation due to the classification of responding units in groups is removed. In other words, one might consider overly restrictive the fact that in the estimation of q_k some units contribute with weight 1 and some others with weight 0. With method (2), the estimate is given by:

$$\hat{q}_k = \sum_{j \in r} D^*(x_k - x_j) / \sum_{j \in s} D^*(x_k - x_j) \quad (2)$$

where D^* has to be specified. In this case, each value x_j contributes towards the estimate \hat{q}_k through D^* , an amount inversely related to the difference $|x_k - x_j|$.

In (2), the problem is twofold: i) to specify the functional form D^* and ii) to define the values of its parameters. In this paper we adopt a function D^* of the normal type:

$$D^*(z) = (h^2 2\pi)^{-1/2} \exp(-z^2/2h^2); \quad z = x_k - x_j, \tag{3}$$

in which the standard deviation, indicated by h , plays a role analogous to that of the parameter h in the expression (1). In both (1) and (2), when h increases, \hat{q}_k approaches to the constant value n_r/n . In (1), it reaches n_r/n when h covers the whole range of the x -values.

An empirical study was designed to evaluate the properties of the proposed procedures, using a very wide range of h values. In the present paper we have limited ourselves to reporting results for only three (constant) values of h , equal to 1/10, 3/10 and 5/10 of the range of the x -sample values. Finally, we must observe that both expressions (1) and (2), apart from a normalizing factor, show themselves as the ratio of two probability density kernel estimators (in the approach of Rosenblatt (1956)) over different sets of x -values. Therefore, as suggested by Giommi (1985b), the value of h may be selected considering proposals put forward in that theory.

3. SUPERPOPULATION MODEL AND ESTIMATORS

For the choice of the estimator of \bar{Y} , we assume a superpopulation model Φ in which the population values $y_k, k=1, 2, \dots, N$, are considered to be a random sample such that:

$$\begin{aligned} E_{\Phi}(Y_k) &= \mu_k = \beta x_k, \\ \text{Var}_{\Phi}(Y_k) &= \sigma_k^2 = \sigma^2 x_k, \end{aligned} \tag{4}$$

where β and Φ unknown and x_k is the known value of the auxiliary variable X . It is apparent that the superpopulation model employed here is mainly applicable to quantitative rather than qualitative variables; other models should be employed in such cases. We further limit ourselves to the consideration of simple random samples. Providing the variance of Y may be specified as in (4), Cassel *et al.* (1983) have shown that the following estimator:

$$T = \bar{X} \left(\sum_r y_k / q_k \right) / \left(\sum_r x_k / q_k \right),$$

where Σ_r indicates the sum over the set r and $\bar{X} = \Sigma_k^N x_k / N$, is approximately unbiased, thanks to the q_k correction, even if the first equation in (4) fails to specify the true relationship between X and Y . This may happen, for example, when the "true" model has an intercept or has two regression coefficients (see (5) below), etc.

Unfortunately, in practice the estimator T cannot be used since q_k is unknown. The problem is, therefore, to evaluate its properties when q_k is replaced by its estimate derived either from method (1) or (2).

We shall examine such estimators, for the three chosen values of h . We denote the estimators by TD_i and TD_i^* where $i=1, 3, 5$ as in Table 1.

Table 1
Definition of Estimators

<i>h</i>	Estimators	
	Method (1)	Method (2)
0.1	TD_1	TD_1^*
0.3	TD_3	TD_3^*
0.5	TD_5	TD_5^*

In addition, also the following estimators are considered in the Monte Carlo study:

$$TC = \bar{X}\left(\sum_s y_k / \sum_s x_k\right) \quad \text{and} \quad TI = \bar{X}\left(\sum_r y_k / \sum_r x_k\right).$$

TC is the full sample estimator, that is, the ratio estimator under the hypothesis of complete response and *TI* is the same estimator based on the set of respondents, on which no *q_k* correction is made for nonresponse. Note that *TI* is also an estimator derived from a well known procedure of imputation (by regression) of missing values (Cassel *et al.* 1983) and equals *TD* when *h* covers the whole range of the *x*-values. *TI* is approximately unbiased only if (4) is true. The bias, as we shall see, depends on the divergence between the condition in (4) and those of the population under study. As in the experiment of the next section model (4) will be a “false” model (that is, the study populations are specified by models different from (4)), the simulation also contributes to the knowledge of this very simple and widely used imputation method.

4. THE MONTE CARLO EXPERIMENT

In the Monte Carlo experiment two populations, POP1 and POP2, were generated following the same procedure as that of Särndal and Hui (1981). POP1 and POP2 are both composed of two strata, say *S1* and *S2*, 500 units each and satisfy the following equations:

$$E_{\Phi}(Y_k) = \beta_1 x_{k1} + \beta_2 x_{k2},$$

$$\text{Var}_{\Phi}(Y_k) = \sigma_1^2 x_{k1} + \sigma_2^2 x_{k2},$$

where $x_{k1} = x_k \partial_k$ and $x_{k2} = x_k (1 - \partial_k)$, with $\partial_k = 1$ if $k \in S1$ and $\partial_k = 0$ if $k \in S2$. The difference between (4) and (5) simulates one of the many errors which one can incur in specifying the superpopulation model. The numerical characteristics of POP1 and POP2 are shown in Table 2.

The simulation procedure can briefly be described in the following steps:

- 1) A simple random sample *s* of *n* (*n* = 50, 100) units is selected from each population:

Table 2
Characteristics of Simulated Populations

Population and strata		POP1				POP2			
		Mean	SD	CV	SK	Mean	SD	CV	SK
Stratum 1	<i>x</i>	19.305	12.71	.66	1.30	20.037	14.50	.72	2.00
	<i>y</i>	7.612	5.38	.71	1.62	1.961	2.21	1.13	3.00
Stratum 2	<i>x</i>	50.325	21.32	.42	.77	49.775	23.28	.47	1.00
	<i>y</i>	30.325	13.38	.44	.72	44.862	21.31	.47	1.00
Total	<i>x</i>	34.815	23.42	.67	.90	34.906	24.44	.70	1.00
	<i>y</i>	18.969	15.26	.80	1.06	23.411	26.25	1.12	1.00

SD = population standard deviation; SK = skewness (3rd moment / (2nd moment)^{3/2}); CV = coefficient of variation.

2) The full sample values are recorded and nonresponse is then generated by each of the two following parametric models:

$$\text{Model A: } q_k = \exp(-\Theta x_k),$$

$$\text{Model B: } q_k = \Theta_1^{\partial_k} \Theta_2^{1-\partial_k}; \quad \partial_k = 1 \text{ (0) if } k \in S1 \text{ (} S2 \text{)},$$

where the parameters Θ , Θ_1 , Θ_2 are chosen in such a way that the average response rate \bar{q} over the whole population is alternatively 0.6 and 0.7. In practice, sets of respondents are obtained by performing a Bernoulli trial for each unit $k \in s$, with probability q_k for "success" (response) and $1 - q_k$ for "failure" (nonresponse).

3) The IRP is estimated by method (1) and (2) and, for each sample, the values of TC , TI , TD , TD^* are calculated.

4) Steps 1 to 3 are repeated 1000 times and at the end we calculate: bias, variance (VAR) and mean squared error (MSE) of the estimators for each sample size (50, 100), response model (A, B), average response rate (0.6, 0.7) and population (POP1, POP2).

The experimental results are reported in Tables 3 and 4.

5. RESULTS OF THE MONTE CARLO EXPERIMENT

Some interesting elements emerge from the examination of Tables 3 and 4.

1. As expected, TC is approximately unbiased in all of the experimental trials.
2. In this experiment the bias of TI is always larger than that of TD and TD^* . Therefore, at least in the situations of the experiment, the adjusted estimator is to be preferred over the non-adjusted one, which corresponds to a procedure of imputation by regression.
3. For the same h value, the bias of TD is always smaller than that of TD^* . The differences are negligible for $h = .1$. As h increases, TD^* tends toward TI faster than TD ; for $h = .5$ the differences between TD^* and TI are irrelevant for practical purposes.
4. The reduction of the bias we are able to obtain using TD instead of TI is always significant, varying from 55% to 82% for model A, from 67% to 92% for model B. TD^* also experiences a notable reduction of the bias: from 51% to 68% for model A, from 61% to 84% for model B.
5. TD and TD^* are equivalent in terms of MSE for $h = .1$, even though TD_1^* is slightly more stable (i.e. has a lower variance). For $h = .3$ and $h = .5$, the lesser stability of TD in comparison with TD^* is generally compensated by the smaller bias, more than enough to make TD preferable to TD^* in terms of MSE.
6. The estimators adjusted by the estimated IRP are not very stable but, in terms of MSE, must be preferred to TI .
7. As expected, the bias is directly related to the increase of the nonresponse rate and to the divergence between the true superpopulation model and the one assumed (i.e. the false model on which the estimators are based). No relevant differences are revealed due to the response models considered in this paper (see Giommi (1984) for the effect of alternative models).
8. The increase of the sample size seems to reduce the bias slightly for all the estimators considered. TD_1 and TD_1^* are exceptions: in this case, the reduction of the bias cannot be attributed to experimental fluctuations but to the actual improvement of the estimate q_k when n increases.

In the end, we may conclude that, in situations similar to the ones considered in this paper, the two methods suggested can be used, with a certain preference for method (1) given its simpler application. The problem of determination of the best value for h (or h_k , in the general case) remains to be examined. We found that, within certain limits, small values for h reduce the bias but also reduce the stability of the adjusted estimator. We have found that, for our experimental examination, the optimum value of h is in the neighbourhood of 0.1. Results obtained from the same experiment but not reported in this paper indicate that a further reduction of h tends to increase the bias. This is to be expected since making h get closer to 0 results in a collection of estimates \hat{q}_k ($k = 1, \dots, n$), equal to 1 and 0 respectively for the respondents and nonrespondents.

6. ACKNOWLEDGEMENT

I am indebted to Prof. Luigi Biggeri for his support throughout the course of the study. I also wish to thank the referees for their helpful comments on the first draft of this paper.

Table 3
Performance of Different Estimators under Response Model A

Estimators		TC	TI	TD ₁	TD ₃	TD ₅	TD ₁ [*]	TD ₃ [*]	TD ₅ [*]
Average response rate $\bar{q} = .60$									
POP1									
$n = 50$	BIAS	.015	.861	.349	.420	.669	.380	.620	.765
	VAR	.405	.973	1.115	1.036	1.007	1.041	.995	.989
	MSE	.405	1.714	1.237	1.212	1.455	1.185	1.379	1.574
$n = 100$	BIAS	.007	.805	.164	.323	.610	.227	.544	.686
	VAR	.186	.416	.443	.429	.412	.415	.404	.402
	MSE	.186	1.064	.470	.533	.784	.467	.700	.873
POP2									
$n = 50$	BIAS	.090	3.125	1.433	1.682	2.544	1.544	2.378	2.887
	VAR	3.952	8.744	9.821	9.823	9.743	9.390	9.233	9.118
	MSE	3.960	18.510	11.874	12.652	16.215	11.774	14.888	17.453
$n = 100$	BIAS	.056	2.959	.749	1.387	2.337	1.004	2.104	2.566
	VAR	1.710	4.144	4.515	5.122	4.819	4.238	4.632	4.518
	MSE	1.713	12.900	5.076	7.046	10.281	5.246	9.059	11.102
Average response rate $\bar{q} = .70$									
POP1									
$n = 50$	BIAS	.015	.581	.226	.271	.418	.249	.415	.439
	VAR	.405	.765	.794	.750	.738	.754	.752	.753
	MSE	.405	1.103	.845	.823	.913	.816	.924	.946
$n = 100$	BIAS	.007	.531	.099	.205	.396	.143	.357	.457
	VAR	.186	.328	.323	.307	.327	.313	.327	.336
	MSE	.186	.610	.333	.349	.484	.333	.454	.544
POP2									
$n = 50$	BIAS	.090	2.130	.813	.939	1.542	.887	1.453	1.822
	VAR	3.952	6.996	7.122	6.827	6.991	6.708	6.753	6.871
	MSE	3.960	11.533	7.783	7.709	9.396	7.495	8.864	10.191
$n = 100$	BIAS	.056	1.966	.473	.953	1.541	.658	1.406	1.731
	VAR	1.710	3.071	3.005	3.062	3.027	2.926	3.008	3.041
	MSE	1.713	6.937	3.229	3.970	5.402	3.359	4.985	6.041

Table 4
Performance of Different Estimators under Response Model B

Estimators		TC	TI	TD ₁	TD ₃	TD ₅	TD ₁ [*]	TD ₃ [*]	TD ₅ [*]
Average response rate $\bar{q} = .60$									
POP1									
n = 50	BIAS	.015	1.086	.290	.383	.716	.323	.688	.992
	VAR	.405	.966	1.208	1.011	.937	1.050	.907	.928
	MSE	.405	2.145	1.29	1.158	1.450	1.154	1.380	1.912
n = 100	BIAS	.007	1.079	.120	.349	.732	.196	.668	.902
	VAR	.186	.422	.513	.429	.420	.447	.401	.403
	MSE	.186	1.586	.527	.551	.956	.485	.847	1.217
POP2									
n = 50	BIAS	.090	4.046	1.362	1.757	2.826	1.562	2.749	3.562
	VAR	3.952	10.285	12.519	12.089	12.010	11.605	11.046	10.994
	MSE	3.960	26.655	14.374	15.176	19.996	14.045	18.603	23.682
n = 100	BIAS	.056	3.897	.454	1.531	2.707	.853	2.521	3.284
	VAR	1.710	4.151	5.432	5.121	5.103	4.798	4.541	4.381
	MSE	1.713	19.338	5.638	7.465	12.431	5.525	10.896	15.166
Average response rate $\bar{q} = .70$									
POP1									
n = 50	BIAS	.015	.584	.179	.221	.409	.196	.376	.499
	VAR	.405	.751	.826	.425	.716	.769	.723	.743
	MSE	.405	1.092	.858	.474	.883	.807	.864	.992
n = 100	BIAS	.007	.536	.046	.173	.365	.087	.317	.436
	VAR	.186	.307	.318	.295	.295	.299	.295	.302
	MSE	.186	.594	.320	.325	.428	.307	.395	.492
POP2									
n = 50	BIAS	.090	2.057	.682	.891	1.477	.804	1.392	1.822
	VAR	3.952	6.199	6.788	6.165	6.232	6.340	6.093	6.270
	MSE	3.960	10.430	7.253	6.959	8.414	6.986	8.031	9.590
n = 100	BIAS	.056	1.918	.157	.755	1.311	.374	1.175	1.562
	VAR	1.710	2.826	2.897	2.884	2.867	2.796	2.836	2.923
	MSE	1.713	6.506	2.922	3.454	4.586	2.936	4.217	5.363

REFERENCES

CASSEL, C.M., SÄRNDAL, C.E., and WRETMAN, J.H. (1983). Some uses of statistical models in connection with the nonresponse problem. In *Incomplete Data in Sample Surveys* (eds. W.G. Madow and I. Olkin), Vol. 3, New York: Academic Press, 143-160.

GIOMMI, A. (1984). On a simple method for estimating individual response probabilities in sampling from finite populations, *Metron*, 42, 185-200.

GIOMMI, A. (1985a). On estimation in nonresponse situations. *Statistica*, 1, 57-63.

GIOMMI, A. (1985b). On the estimation of the individual response probabilities. *Proceedings of the 45th Session of the International Statistical Institute*, Vol. 2 (Contributed Papers), 577-578.

- ROSENBLATT, M. (1956). Remarks on some nonparametric estimates for the density function. *Annals of Mathematical Statistics*, 27, 832-837.
- SÄRNDAL, C.E., and HUI, T.K., (1981). Estimation for nonresponse situations: to what extent must we rely on models? In *Current Topics in Survey Sampling*, (eds. D. Krewski, R. Platek and J.N.K. Rao), New York: Academic Press, 227-246.

Estimates Based on Randomly Rounded Data

C.S. WITHERS¹

ABSTRACT

Methods are given to estimate functions of the cell probabilities associated with a table of multinomial data that has been randomly rounded to multiples of a given number, say l . We show that: (i) random rounding causes only second order effects on bias and variance; (ii) the loss of efficiency in using the natural estimates of cell probability is negligible provided that the cell entry is large compared with $(j^2 - 1) / (6R)$ where R is the number of cells in the table; and (iii) estimates of apparently exponentially small bias are available for moments of these natural estimates and for polynomials in the cell probabilities.

KEY WORDS: Random rounding; Bias reduction; Efficiency.

1. INTRODUCTION AND SUMMARY

This paper gives methods of estimating a function of the cell probabilities associated with a table of multinomial data that has been randomly rounded. Random rounding is a widely used method for preserving confidentiality in situations where an entry of 1 in a table might identify an individual and so break a confidentiality requirement. Instead of tabling the value of a table entry, say N , one rounds N to the nearest multiple of a given number l above or below N with probability (w.p.) α or below N w.p. $1 - \alpha$, where α is chosen so that the rounded value M satisfies

$$E(M | N) = N.$$

That is, if for some integer j , $jl \leq N < (j + 1)l$, then

$$M = \begin{cases} jl \text{ w.p. } 1 - \alpha \\ (j + 1)l \text{ w.p. } \alpha \end{cases} \quad (1.1)$$

where $\alpha = r/l$ and $r = N - jl$.

The rounding base l used by the Department of Statistics in New Zealand is $l = 3$, while Statistics Canada reportedly uses $l = 5$. See Penny and Ryan (1986).

Random rounding should not be confused with grouping or non-random rounding of sample values to the nearest integral multiple of l (associated with Sheppard's corrections for moments). Nor should it be confused with intentional contamination, another method of preserving confidentiality where one simply adds to N an independent random variable with mean 0. (The main disadvantage of intentional contamination is the possibility of a negative cell entry). For some references on these methods see Gastwirth *et al.* (1978) and Kendall and Stuart (1977). Some references on random rounding for multivariate data and grouped data are also given in Gastwirth *et al.* (1978).

¹C.S. Withers, Applied Mathematics Division, Department of Scientific and Industrial Research, Box 1335, Wellington, New Zealand.

In this paper we confine our attention to problems of estimating a function of the cell probabilities associated with a table of R values that have been randomly rounded. For convenience we label these cell probabilities as p_1, \dots, p_R rather than $\{p_{ij}, 1 \leq i \leq I, 1 \leq j \leq J\}$, as is more usual for an $I \times J$ table.

Thus, $1 = \sum_1^R p_i$ and $n = \sum_1^R N_i$ is the sum of the entries in the table. Let $\{M_i\}$ be the rounded values of $\{N_i\}$. Given n , we assume $\{N_i\}$ has the multinomial distribution with parameters n and $\{p_i\}$. This is true with $p_i = m_i / \sum_j m_j$ if, unconditionally, $\{N_i\}$ are independent Poisson variables with means $\{m_i\}$.

Two unbiased estimates of p_1 are

$$p_1^* = N_1/n \text{ and } \hat{p}_1 = M_1/n. \quad (1.2)$$

The first is not a true estimate since N_1 is not made available. The second is the natural estimate. (We assume n is reported. If it is not, there is negligible difference in replacing n by $\sum_1^R M_i$.) However, other unbiased estimates exist, namely the "complementary estimate"

$$\tilde{p}_1 = - \sum_{j \neq 1} M_j/n, \quad (1.3)$$

and hence

$$p_1(\lambda) = (1 - \lambda)\hat{p}_1 + \lambda\tilde{p}_1 \text{ for any given } \lambda. \quad (1.4)$$

This raises the issue of what is the best λ to use, and what loss of efficiency there is in sticking to the natural estimate — that is, using $\lambda = 0$. An answer requires the variances of these estimators. These are given by

Theorem 1.1.

$$\text{var}(\hat{p}_1) = (p_1 - p_1^2) n^{-1} + \{(l^2 - 1)/6 + \Delta_n(p_1)\}n^{-2} = v_n(p_1), \quad (1.5)$$

where

$$\Delta_n(p_1) = \sum_{i=0}^{l-1} i(l-i) \{P(N_1 \bmod l = i) - l^{-1}\}. \quad (1.6)$$

Also,

$$\text{var}(\tilde{p}_1) = (p_1 - p_1^2)n^{-1} + \{(R-1)(l^2-1)/6 + \sum_{j \neq 1} \Delta_n(p_j)\}n^{-2}, \quad (1.7)$$

and

$$\text{var}(p_1(\lambda)) = (p_1 - p_1^2)n^{-1} + \{\alpha(\lambda)(l^2-1)/6 + \nabla_n(p)\}n^{-2}, \quad (1.8)$$

where

$$\alpha(\lambda) = (1 - \lambda)^2 + (R-1)\lambda^2 \quad (1.9)$$

and

$$\nabla_n(p) = (1 - \lambda)^2 \Delta_n(p_1) + \lambda^2 \sum_{i \neq 1} \Delta_n(p_i). \quad (1.10)$$

Proofs of the theorems in this paper are given in Section 2.

In Appendix A we give evidence that for $0 < p_1 < 1$, $P(N_1 \bmod l = i) - l^{-1} \rightarrow 0$ exponentially fast as $n \rightarrow \infty$, so that $\Delta_n(p_1) \rightarrow 0$ exponentially fast as $n \rightarrow \infty$, and hence $\nabla_n(p)$ also, provided $p_i \neq 0$ for all i .

Since $\alpha(\lambda)$ is minimised by $\lambda_R = R^{-1}$ and $\alpha(\lambda_R) = 1 - R^{-1}$ so, asymptotically, is $\text{var}(p_1(\lambda))$. Hence the loss of efficiency in using the natural estimate \hat{p}_1 rather than the asymptotically optimal unbiased estimate $p_1(\lambda_R)$ when R is large, is

$$\{\text{var}(\hat{p}_1) - \text{var}(p_1(\lambda_R))\} / \text{var}(p_1(\lambda_R)) \approx (l^2 - 1) / \{6Rn(p_1 - p_1^2)\} \quad (1.11)$$

which is negligible provided $M_1(1 - M_1/n) \approx n(p_1 - p_1^2)$ is large compared with $(l^2 - 1) / \{6R\}$.

Generally $M_1(1 - M_1/n)$ can be approximated by M_1 . This then gives a convenient rule of thumb as to when the natural estimates are efficient. (If one or more $\{p_i\}$ are zero, since $p_i = 0$ implies $N_i = M_i = 0$, $\Sigma_{i \neq 1}$ must be interpreted as excluding cells for which $p_i = 0$, and R as the number of cells in the table for which $p_i \neq 0$.)

Using (1.5) we can now make a brief comparison with the method of contamination. The Australian and U.K. statistics departments reportedly round by adding to each cell entry 1 w.p. 1/4, 0 w.p. 1/2 and -1 w.p. 1/4, so that

$$\text{var}(\hat{p}_1) = (p_1 - p_1^2)n^{-1} + 1/2n^{-2}.$$

The factor 1/2 improves on 4/3 for the New Zealand system ($l = 3$) and 4 for the Canadian system ($l = 5$). The cost is less protection (a maximum change of 1 as opposed to 2 for the New Zealand system and 4 for the Canadian system), and a possibly negative cell entry if the procedure is applied to cells with zero entries.

Theorem 1.1 shows that random rounding has only a second order effect on the efficiency of estimating p_1 — the variance is only increased by a term of magnitude n^{-2} . The next result shows that this very important result is also true for estimating any smooth function of $\{p_i\}$. Set $r = R - 1$, $\mathbf{p} = (p_1, \dots, p_r)$, $\mathbf{N} = (N_1, \dots, N_r)$, $\mathbf{M} = (M_1, \dots, M_r)$, $\mathbf{p}^* = \mathbf{N}/n$ and $\hat{\mathbf{p}} = \mathbf{M}/n$. Thus we have $\text{cov}(\mathbf{p}^*) = V/n$ where $V = \text{diag}(\mathbf{p} - \mathbf{p}\mathbf{p}')$. Suppose now we wish to estimate $f(\mathbf{p})$, a function with continuous second derivatives.

That is, $\dot{f}(\mathbf{p}) = \partial f(\mathbf{p}) / \partial \mathbf{p}$ is a continuous $r \times 1$ function and $\ddot{f}(\mathbf{p}) = \partial^2 f(\mathbf{p}) / \partial \mathbf{p} \partial \mathbf{p}'$ is a continuous $r \times r$ function.

Theorem 1.2. As $n \rightarrow \infty$ both $E(f(\mathbf{p}^*))$ and $E(f(\hat{\mathbf{p}}))$ equal

$$f(\mathbf{p}) + B(\mathbf{p})n^{-1} + O(n^{-2}) \text{ where } B(\mathbf{p}) = \text{trace}(\ddot{f}(\mathbf{p})V/2). \quad (1.12)$$

Also both $\text{var}(f(\mathbf{p}^*))$ and $\text{var}(f(\hat{\mathbf{p}}))$ equal

$$v(\mathbf{p})n^{-1} + O(n^{-2}) \text{ where } v(\mathbf{p}) = \dot{f}(\mathbf{p})' V \dot{f}(\mathbf{p}). \quad (1.13)$$

This theorem shows that

- (a) random-rounding increases the variance of the natural estimate for $f(\mathbf{p})$ by only $O(n^{-2})$; and
- (b) random-rounding likewise has only a second order effect on the bias of the natural estimate for $f(\mathbf{p})$.

According to (1.12), the natural estimate of $f(\mathbf{p})$, $f(\hat{\mathbf{p}})$, has bias of magnitude n^{-1} . We now show how to reduce this to n^{-2} .

Corollary 1.1. If for some function $f_n(\mathbf{p})$, $E(f_n(\mathbf{p}^*)) = f(\mathbf{p}) + O(n^{-2})$ then $E(f_n(\hat{\mathbf{p}})) = f(\mathbf{p}) + O(n^{-2})$.

Two such choices for $f_n(\hat{\mathbf{p}})$ are the "delta-estimate" for which

$$f_n(\mathbf{p}) = f(\mathbf{p}) - \left\{ \sum_{i=1}^r f_{ii}(\mathbf{p}) p_i - \mathbf{p}' \ddot{f}(\mathbf{p}) \mathbf{p} \right\} / (2n), \quad (1.14)$$

where $f_{ii}(\mathbf{p}) = \partial^2 f(\mathbf{p}) / \partial p_i^2$, and the "jack-knife estimate" for which

$$f_n(\mathbf{p}) = n f(\mathbf{p}) - (n-1) \bar{f}, \quad (1.15)$$

where

$$\bar{f} = \sum_{i=1}^r p_i f([\mathbf{n}\mathbf{p} - \mathbf{e}_i] / (n-1)) + (1 - \sum_{i=1}^r p_i) f([\mathbf{n}\mathbf{p} / (n-1)]),$$

\mathbf{e}_i = the i -th unit vector in R^r ,

$$\text{and } [x] : R^r \rightarrow R^r \text{ is defined by } [x]_i = \begin{cases} 0, & x_i < 0 \\ x_i, & 0 \leq x_i \leq 1 \\ 1, & x_i > 1 \end{cases}$$

These estimates were derived in Withers (1987a and 1987b). In particular, if $f(\mathbf{p})$ is only a function of p_1 , say $f(\mathbf{p}) = g(p_1)$, then $f_n(\mathbf{p}) = g(p_1) - \ddot{g}(p_1)(p_1 - p_1^2) / (2n)$ and $\bar{f} = p_1 g([\mathbf{n}p_1 - 1] / (n-1)) + (1 - p_1) g([\mathbf{n}p_1 / (n-1)])$. For example if $f(\mathbf{p}) = p_1^2$ then the delta-estimate uses $f_n(\mathbf{p}) = p_1^2 \{1 - (1 - p_1)/n\}$.

We now illustrate that if $f(\mathbf{p})$ is a polynomial we can in fact find an estimate of $f(\mathbf{p})$ based on the natural estimate with bias apparently exponentially small. We do this for the case $f(\mathbf{p}) = p_1^2$.

Theorem 1.3. $\hat{\lambda}_1 = \{\hat{p}_1^2 - n^{-1}\hat{p}_1 - n^{-2}(l^2 - 1)/6\} (1 - n^{-1})^{-1}$ estimates $\lambda_1 = p_1^2$ with bias $\Delta_n(p_1)(n^2 - n)^{-1}$.

Similarly if $f_n(\mathbf{p})$ is a moment of $\hat{\mathbf{p}}$ then we can also find an estimate of $f_n(\mathbf{p})$ with bias apparently exponentially small. We illustrate this for the case $f_n(\mathbf{p}) = \text{var}(\hat{p}_1)$.

Theorem 1.4. $\hat{\lambda}_{2n} = n^{-1}(\hat{p}_1 - \hat{\lambda}_1) - n^{-2}(l^2 - 1)/6$ estimates $\lambda_{2n} = \text{var}(\hat{p}_1)$ with bias $-\Delta_n(p_1)(n^2 - n)^{-1}$.

These results may be generalised to higher order polynomials and moments using the expression for moments and cumulants of $\hat{\mathbf{p}}$ given in Appendix B. We now show that for the special case of $f(\mathbf{p})$ collinear, an unbiased estimate exists.

Theorem 1.5. Set $f_I(\mathbf{p}) = \Pi_{i=1}^I p_i$ where $1 \leq I \leq R$ and

$$a_{nI} = n^{-I} n! / (n - I)! = (1 - n^{-1})(1 - 2n^{-1}) \dots (1 - \{I - 1\}n^{-1}). \quad (1.16)$$

Then

$$E(f_I(\hat{\mathbf{p}})) = E(f_I(\mathbf{p}^*)) = f_I(\mathbf{p}) a_{nI}. \quad (1.17)$$

Hence an unbiased estimate of $f(\mathbf{p})$ is $f_I(\hat{\mathbf{p}}) / a_{nI}$.

Corollary 1.2. $\text{cov}(\hat{p}_1, \hat{p}_2) = -p_1 p_2 / n$. Its unbiased estimate is $-\hat{p}_1 \hat{p}_2 / (n - 1)$. More generally for $1 \leq I \leq R$, $E(\Pi_{i=1}^I (\hat{p}_i - p_i)) = c_{nI} \Pi_{i=1}^I p_i$ with unbiased estimate $(\Pi_{i=1}^I \hat{p}_i) a_{nI} / c_{nI}$ where $c_{nI} = \sum_{j=0}^I (-1)^{I-j} \binom{I}{j} a_{nj}$. (The same result holds with $\hat{\mathbf{p}}$ replaced by \mathbf{p}^* .)

From (1.16) one may derive unbiased estimates for other special polynomials in \mathbf{p} such as p_1^2 , $p_1 p_2 (p_1 + p_2)$ and $\sum_{i=1}^R p_i^3$ - but not for $p_1^2 p_2$ or p_1^3 .

Corollary 1.3. For $1 \leq I < R$ an unbiased estimate of

$$f_I(\mathbf{p}) \sum_1^I p_i \text{ is } f_I(\hat{\mathbf{p}}) \left\{ 1 - In^{-1} - \sum_{I+1}^R \hat{p}_i \right\} / a_{n, I+1}. \quad (1.18)$$

In particular an unbiased estimate of p_1^2 is

$$\hat{p}_1 (\bar{p}_1 - n^{-1}) (1 - n^{-1})^{-1}. \quad (1.19)$$

We emphasize that the results of this paper are based on the assumption that table entries are independent Poisson's, or at least multinomial conditional on the total. The Poisson and multinomial models are appealing as they have a ready interpretation, and because sums of Poisson variables are Poisson. But sums of multinomials are multinomial only if they share the same cell probabilities \mathbf{p} . This suggests that conclusions drawn from such models may be less accurate if the populations modelled are composed of two or more inhomogeneous groups.

2. PROOFS

Proof of Theorem 2.1. Set $r = N_1 \bmod l$. Then (1.1) holds for $N = N_1$, $M = M_1$ with $il = N - r$ and

$$E(M_1^2 | r) = (N_1 - r)^2 (1 - r/l) + (N_1 - r + l)^2 r/l = N_1^2 + lr - r^2.$$

Hence

$$E(\hat{p}_1^2) = E(p_1^{*2}) + n^{-2} A_n(p_1), \quad (2.1)$$

where

$$\begin{aligned} A_n(p_1) &= E(M_1^2 - N_1^2) = E(lr - r^2) = \sum_{i=0}^{l-1} (li - i^2) P(N = i) \\ &= (l^2 - 1)/6 + \Delta_n(p_1) \end{aligned}$$

since

$$l^{-1} \sum_{i=0}^{l-1} i(l - i) = (l^2 - 1)/6. \quad (2.2)$$

But

$$E(p_1^{*2}) = p_1^2 + (p_1 - p_1^2)n^{-1}, \quad (2.3)$$

so (1.5) follows. Now $\tilde{p}_1 = \hat{p}_1 - \sum (M_j - N_j) / n$,

$$\begin{aligned} \text{so } E(\tilde{p}_1^2) &= E(\hat{p}_1^2) - 2n^{-2} \sum E(M_1(M_j - N_j)) + n^{-2} \sum E((M_i - N_i)(M_j - N_j)) \\ &= E(\hat{p}_1^2) - 2n^{-2} A_n(p_1) + n^{-2} \sum A_n(p_i) \end{aligned}$$

$$\text{since } E(\Pi_i f_i(M_i) | \{N_i\}) = \Pi_i E(f_i(M_i) | N_i). \quad (2.4)$$

Hence $\text{var}(\tilde{p}_1) = (p_1 - p_1^2)n^{-1} + n^{-2} \sum_{i \neq 1} A_n(p_i)$ so (1.7) holds.

Also,

$$\begin{aligned} E(\hat{p}_1 \tilde{p}_1) &= p_1 - n^{-2} \sum_{i \neq 1} E(M_1 M_i) = p_1 - \sum_{i \neq 1} E(p_1^* p_i^*) \\ &= p_1 - \sum_{i \neq 1} p_1 p_i (1 - n^{-1}) = p_1 - p_1 (1 - p_1) (1 - n^{-1}), \end{aligned}$$

so

$$\text{cov}(\hat{p}_1, \tilde{p}_1) = (p_1 - p_1^2)n^{-1}. \quad (2.5)$$

Hence $\text{var}(p_1(\lambda)) = (p_1 - p_1^2)n^{-1} + \{(1 - \lambda)^2 A_n(p_1) + \lambda^2 \sum_{i \neq 1} A_n(p_i)\}n^{-2}$ and (1.8) holds.

Proof of Theorem 1.2. This was proved for \mathbf{p}^* in Withers (1987a). Also since \dot{f} is finite in a neighborhood of \mathbf{p} ,

$$f(\hat{\mathbf{p}}) = f(\mathbf{p}^*) + (\hat{\mathbf{p}} - \mathbf{p}^*)' \dot{f}(\mathbf{p}^*) + O(|\hat{\mathbf{p}} - \mathbf{p}^*|^2).$$

$$E((\hat{\mathbf{p}} - \mathbf{p}^*) | N) = 0, E((\hat{p}_1 - p_i^*)^2 | N) = 2n^{-2} I(N_1 \bmod l \neq 0),$$

where $I(A) = 1$ or 0 for A true or false, that is, $I(\cdot)$ is the indicator function. Hence $E(f(\hat{\mathbf{p}})) = E(f(\mathbf{p}^*)) + O(n^{-2})$ and $\text{var}(f(\hat{\mathbf{p}})) = \text{var}(f(\mathbf{p}^*)) + O(n^{-2})$.

Proof of Theorem 1.3. This follows directly from (2.1) and (2.3).

Proof of Theorem 1.4. This follows from (2.1) and (1.5).

Proof of Theorem 1.5. The first equality in (1.16) follows from (2.4), and the second from the multinomial theorem. Corollary 1.2 follows immediately.

Proof of Corollary 1.3. From (1.16), for $1 \leq I < i \leq R$ we have

$$E(f_I(\hat{\mathbf{p}}) \hat{p}_i) = f_I(\mathbf{p}) p_i a_{n, I+1}$$

so

$$\begin{aligned}
 E(f_I(\hat{\mathbf{p}}) \sum_{I+1}^R \hat{p}_i / a_{n,I+1}) &= f_I(\mathbf{p}) (1 - \Sigma_1^I p_i) \\
 &= E(f_I(\hat{\mathbf{p}}) / a_{nI}) - f_I(\mathbf{p}) \Sigma_1^I p_i.
 \end{aligned}$$

ACKNOWLEDGEMENT

I wish to thank Peter McGavin for doing the computing in Appendix A.

APPENDIX A

One expects that for f a smooth function

$$E(f(\hat{\mathbf{p}})I(N_1 = j_1 \bmod l, \dots, N_s = j_s \bmod l)) \rightarrow f(\mathbf{p})l^{-s} \quad (\text{A.1})$$

as $n \rightarrow \infty$ provided $0 < p_i < 1$ for $1 \leq i \leq s \leq R$.

If $E(f(\hat{\mathbf{p}})) = f(\mathbf{p})$, one expects the rate of convergence to be exponential, $O(e^{-\lambda n})$ for some $\lambda > 0$. If $f(\hat{\mathbf{p}})$ is biased, then its bias is $O(n^{-1})$, so that one would expect this rate also to apply to (A.1). Convergence will in general break down as \mathbf{p} approaches the boundary of $[0, 1]^r$, since

$$\begin{aligned}
 &E(f(\hat{\mathbf{p}})I(N_1 = j_1 \bmod l, \dots, N_s = j_s \bmod l)) \\
 &= \begin{cases} f(\mathbf{p})I(j_1 = j_2 = \dots = j_s = 0) & \text{if } \mathbf{p} = 0 \\ f(\mathbf{p})I(j_1 = n \bmod l) & \text{if } p_1 = 1. \end{cases}
 \end{aligned}$$

To test these expectations we considered the case $s = 1$, $l = 3$, $j = 0$ and the functions (a) $f(\mathbf{p}) = 1$, (b) $f(\mathbf{p}) = p_1$, and (c) $f(\mathbf{p}) = \exp(p_1)$. Computations were done in quadruple precision on a VAX11/780, giving a precision for

$$\Delta = E(f(\hat{\mathbf{p}})I(N_1 = j_1 \bmod l, \dots, N_s = j_s \bmod l)) - f(\mathbf{p})l^{-s}$$

of 112 bits – nearly 34 decimal places. Figures 1a, 1b and 1c plot Δ versus p_1 for $n = 6, 18, 54$. Since $n \bmod 3 = 0$, Δ is symmetric about $p_1 = 1/2$ for (a).

Since $\Delta = 2/3f(0)$ at $p_1 = 0$, and is equal to $2/3$, 0 and $2/3$ for (a), (b) and (c) respectively, convergence breaks down at $p_1 = 0$ for (a) and (c), but not for (b). At $n = 18$, Δ is already negligibly different from 0 for p_1 in $(.2, .8)$ for (a) and for p_1 in $(.1, .8)$ for (b) and (c). At $n = 54$, these ranges have grown to cover $(.1, .9)$ for (a), $(.02, .95)$ for (b), and $(.07, .95)$ for (c).

Figures 2a and 2b plot $Y = \log(-\log|\Delta|)$ versus $X = \log(n)$ for (a) $f(\mathbf{p}) = 1$ and (b) $f(\mathbf{p}) = p_1$. As expected, except for small n , the curves are roughly parallel to $Y = X$ (except for (b) with $p_1 = .01$), consistent with $\Delta = O(e^{-\lambda n})$ for some $\lambda > 0$. The curves are not smooth, as Δ has only been calculated at n a power of 2 ($n = 2^i$ for $0 \leq i \leq 7$).

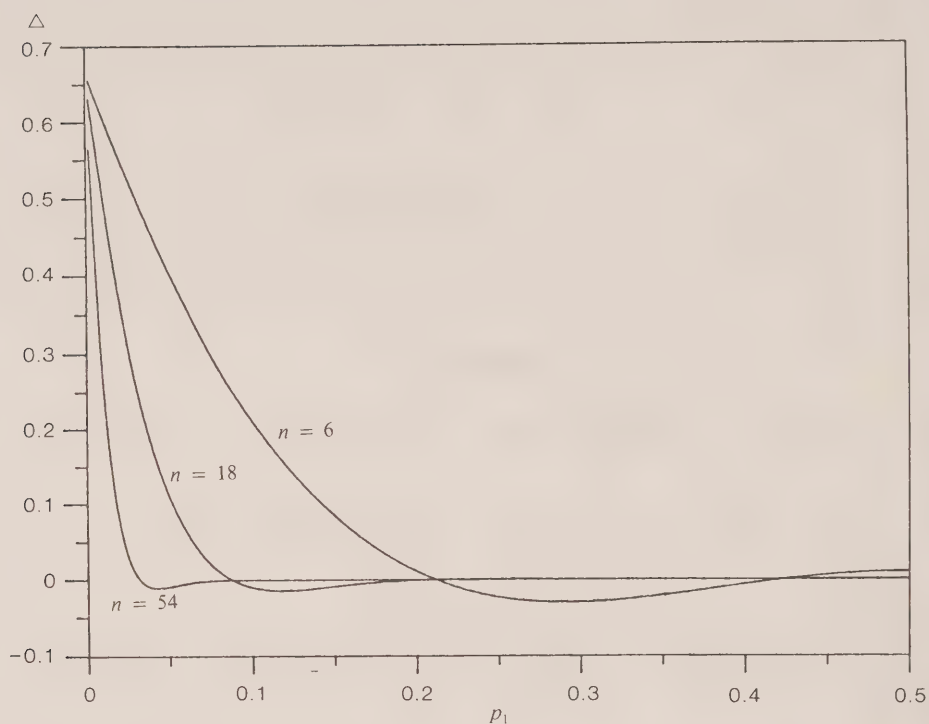


Figure 1a. Evidence for (A.1) When $f(\mathbf{p}) = 1$.

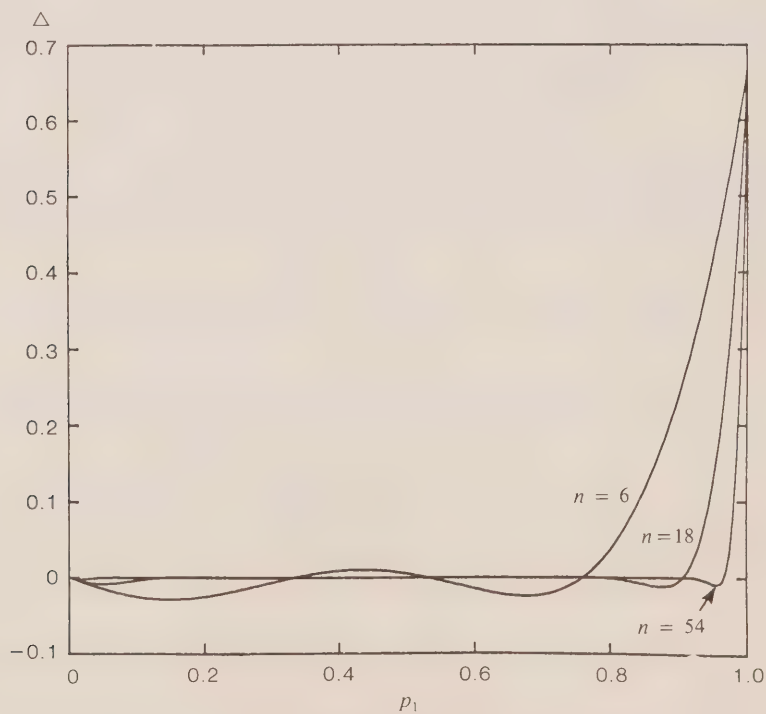


Figure 1b. Evidence for (A.1) When $f(\mathbf{p}) = p_1$.

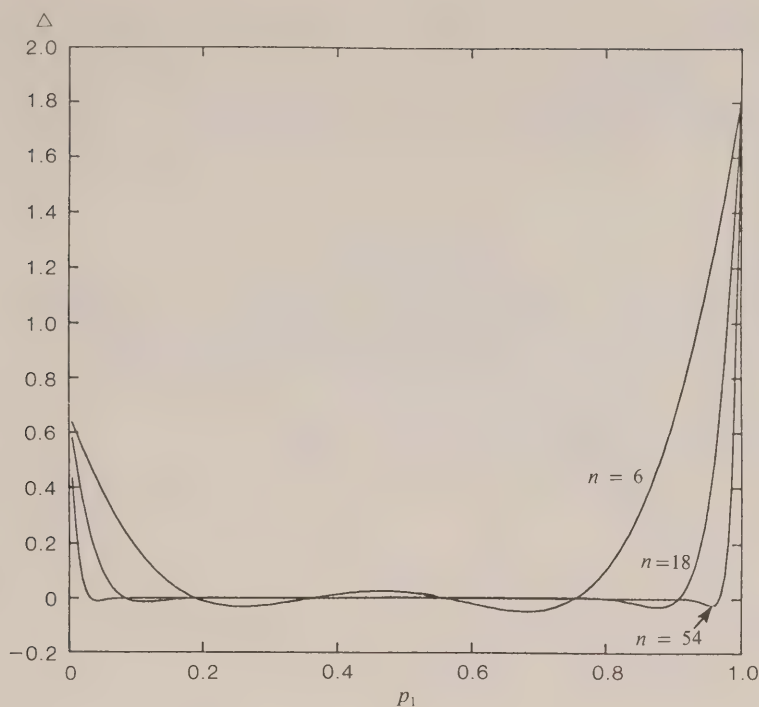


Figure 1c. Evidence for (A.1) When $f(\mathbf{p}) = \exp(p_1)$.

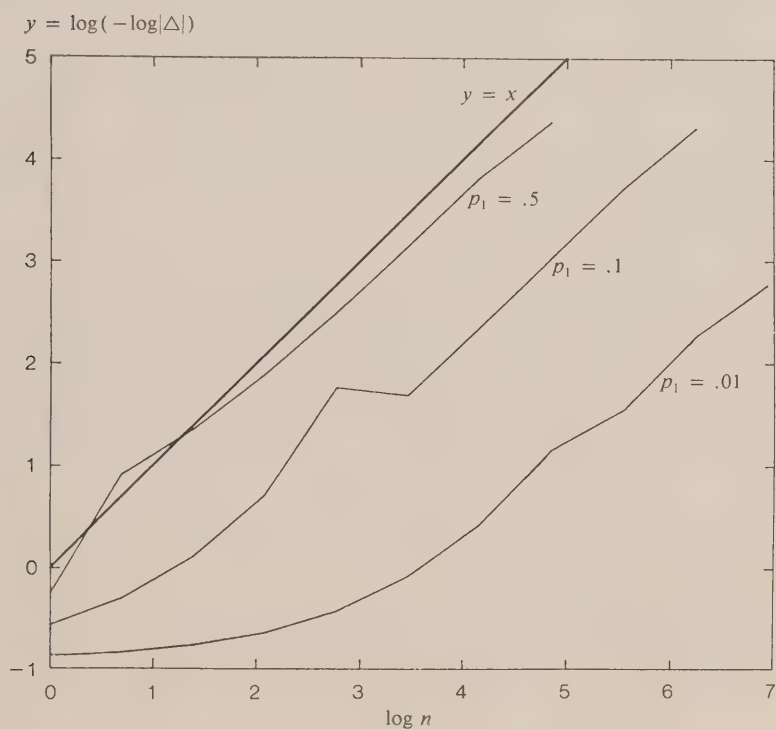


Figure 2a. Evidence for Exponential Convergence in (A.1) for $f(\mathbf{p}) = 1$.

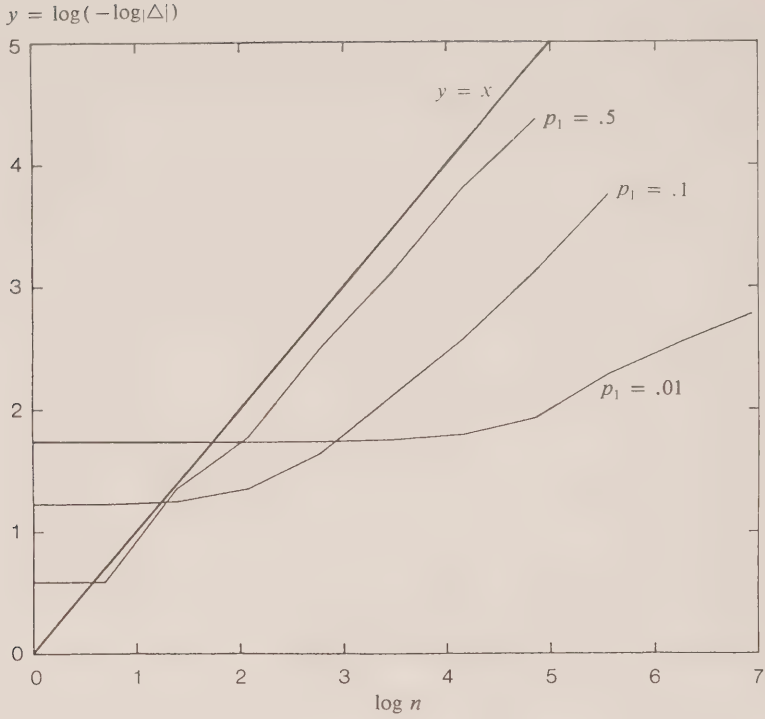


Figure 2b. Evidence for Exponential Convergence in (A.1) for $f(\mathbf{p}) = p_1$.

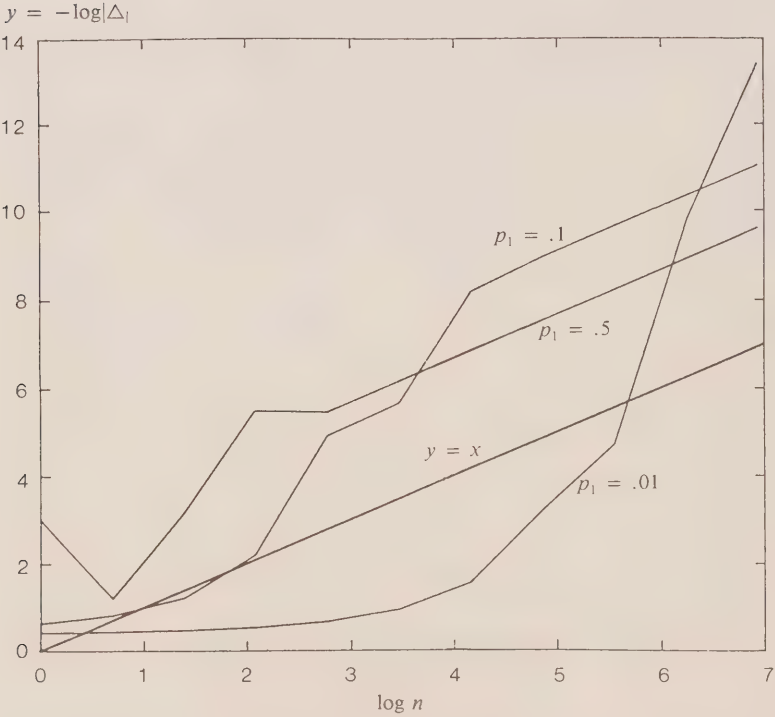


Figure 3. Evidence for Convergence at Rate $\sim n^{-1}$ in (A.1) for $f(\mathbf{p}) = \exp(p_1)$.

Figure 3 plots $Y = -\log |\Delta|$ versus $X = \log(n)$ for (c) $f(\mathbf{p}) = \exp(p_1)$. For n large the curves are parallel to $Y = X$ for $p_1 = .5$ and $.1$ consistent with $\Delta = O(n^{-1})$, but for $p_1 = 0.1$ the increase is much faster than linear. The graphs generally confirm our expectations on the rate of convergence in (A.1). To obtain analytic proofs would appear to require some sophisticated number theory.

APPENDIX B

Here we compare the moments and cumulants of $\mathbf{p}^* = \mathbf{N}/n$ and $\hat{\mathbf{p}} = \mathbf{M}/n$. Set $p_1 = 1 - p_l$, $n_i = N_i \bmod l$, and $m_i(j) = E(p_1^{*j} I(n_1 = j)) \rightarrow p_1^j/l$ as $n \rightarrow \infty$, assuming $p_1 \neq 0$ or 1 . Elementary calculations yield

$$\mu(\hat{\mathbf{p}}) = \mu(\mathbf{p}^*) = \mathbf{p},$$

$$\mu_2(\hat{p}_1) = \mu_2(p_1^*) + M_{22}n^{-2} = p_1q_1n^{-1} + O(n^{-2}),$$

where

$$M_{22} = A_n(p_1) = \sum_{i=0}^{l-1} i(l-i)m_0(i) \rightarrow (l^2 - 1)/6$$

as $n \rightarrow \infty$,

$$\mu_3(\hat{p}_1) = \mu_3(p_1^*) + 3n^{-2} \sum_{j=0}^{l-1} (lj - j^2)\{m_2(j) - 2p_1m_1(j) + p_1^2m_0(j)\}$$

$$+ n^{-3} \sum_{j=0}^{l-1} a_{jl}m_0(j)$$

$$= \mu_3(p_1^*) + o(n^{-2}) = p_1q_1(1 - 2p_1)n^{-2} + o(n^{-2}),$$

and

$$a_{jl} = -j^3(1 - j/l) + (l - j)^3j/l.$$

Similarly $\mu_4(\hat{p}_1)$ has the form $\mu_4(p_1^*) + \sum_2^4 M_{4i}n^{-i} = O(n^{-2})$ and $\kappa_4(\hat{p}_1)$ has the form $k_{4i}n^{-i}$ where $k_{42} = M_{42}$ does not converge to 0 as $n \rightarrow \infty$. Hence $\kappa_4(\hat{p}_1) \sim n^{-2}$, not n^{-3} . Hence $\hat{\mathbf{p}}$ does not satisfy the Cornish-Fisher assumption that $\kappa_r(\hat{\mathbf{p}}) = O(n^{1-r})$ for $r \geq 1$: see for example Kendall and Stuart (1977).

Moments and cumulants may also be obtained from the m.g.f. (moment generating function), which we now obtain.

$$E(\exp(t_1 M_1/n) \mid N_1) = \exp(t_1 N_1/n) S(t_1, n_1)$$

where

$$S(t_1, n_1) = (1 - n_1/l) \exp(-n_1 t_1/n) + (n_1/l) \exp((l - n_1)t_1/n).$$

Hence by (2.4), the m.g.f. is

$$E(\exp(t' \hat{\mathbf{p}})) = E(\exp(t' \mathbf{N}/n)) S(t) \text{ where } S(t) = \prod_1^l S(t_i, n_i).$$

Also at $t = 0$, $S_1 = 0$ and so $S_{ij\dots} = 0$ if a subscript occurs exactly once. For example setting

$$S = S(t), \partial_i = \partial / \partial t_i, S_i = \partial_i S, S_{ij} = \partial_i \partial_j S, \dots$$

gives

$$\begin{aligned} E(\hat{p}_1^2 \exp(t' \hat{\mathbf{p}})) &= E(\exp(t' \mathbf{N} / n) \{p_1^{*2} S + 2p_1^* S_1 + S_{11}\}), \\ E(\hat{p}_1^2 \hat{p}_2^2 \exp(t' \hat{\mathbf{p}})) &= E(\exp(t' \mathbf{N} / n) \{p_1^{*2} (p_2^{*2} S + 2p_2^* S_2 + S_{22}) + \\ &\quad 2p_1^* (p_2^{*2} S_1 + 2p_2^* S_{12} + S_{122}) + (p_2^{*2} S_{11} + 2p_2^* S_{112} + S_{1122})\}) \end{aligned}$$

Hence $E(\hat{p}_1^2) = E\{p_1^{*2} + S_{11}(0)\}$ and

$$E(\hat{p}_1^2 \hat{p}_2^2) = E\{p_1^{*2} p_2^{*2} + p_1^{*2} S_{22}(0) + p_2^{*2} S_{11}(0) + S_{1122}(0)\},$$

where $S_{ii}(0) = S_{11}(0, n_i) = n^{-2}(l - n_i)n_i = n^{-2} \sum_{k=0}^{l-1} (l - k)kI(n_i = k)$ and $S_{1122}(0) = S_{11}(0) S_{22}(0)$. Some further simplifications can be obtained using $N_2 | N_1 \sim Bi(\theta, n - N_1)$ where $\theta = p_2 / (1 - p_1)$. From the multinomial m.g.f. one obtains

$$E(p_1^{*2} p_2^{*2}) = n^{-4} p_1 p_2 \{(n)_4 p_1 p_2 + (n)_3 (p_1 + p_2) + (n)_2\}$$

where $(n)_i = n! / (n - i)! = n(n - 1) \dots (n - i + 1)$.

REFERENCES

- GASTWIRTH, J.L., KRIEGE, A.M., and RUBIN, D.B. (1978). Statistical analyses from summary data and their impact on the issue of confidentiality. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 183-188.
- KENDALL, M.G., and STUART, A. (1977). *The Advanced Theory of Statistics, Volume 7*. London: Griffin.
- NARGUNDAR, M.S., and SAVELAND, W. (1972). Random rounding to prevent statistical disclosure. *Proceedings of the Social Statistics Section, American Statistical Association*, 382-385.
- PENNY, R., and RYAN, M. (1986) A problem associated with random-rounding. *New Zealand Statistician*, 21, 43-52.
- WITHERS, C.S. (1987a). Bias reduction by Taylor series. *Communications in Statistics - Theory and Methods* (forthcoming).
- WITHERS, C.S. (1987b). Jackknifing binomials and multinomials. Unpublished manuscript, Department of Scientific and Industrial Research.

Variance Estimation for the Canadian Labour Force Survey

G.H. CHOUDHRY and H. LEE¹

ABSTRACT

The biases and stabilities of alternative variance estimators for the two stage random group design (Rao et al. 1962) are evaluated in a Monte Carlo study in the context of Canadian Labour Force Survey. The variance formula for raking ratio estimation procedure is derived using Taylor linearization method. The properties of the variance formula are investigated by a Monte Carlo simulation.

KEY WORDS: Keyfitz's variance estimator; Raking ratio estimator; Taylor linearization; Monte Carlo simulation.

1. INTRODUCTION

The Canadian Labour Force Survey (LFS) is the largest monthly household survey conducted by Statistics Canada and is used to produce estimates of various labour force characteristics at national, provincial and sub-provincial levels. It follows a stratified multi-stage rotating sample design with six rotation panels (Platek and Singh 1976).

Following each decennial census of population, the LFS has undergone a sample redesign. As part of the 1981 post-censal redesign, an extensive program of research was undertaken in the areas of sampling, data collection, and estimation methodologies (Singh and Drew 1981). The post-stratified ratio estimation procedure used in the old design was replaced by a raking ratio estimation procedure to improve the reliability of subprovincial data. This paper presents the results related to variance estimation methodology.

The methodology for variance estimation for the old LFS was based on Woodruff's generalization (Woodruff 1971) of the Keyfitz procedure (Keyfitz 1957) using Taylor linearization applied to the post-stratified ratio estimates (Platek and Singh 1976). This method will be called the Keyfitz method as in Platek and Singh (1976).

There are three area types identified in the LFS design, i.e., self-representing (SR) areas consisting of major cities, non-self-representing (NSR) areas which are smaller urbans and rural areas, and special areas composed of military, institutions and remote areas. For the NSR and special areas it was decided to use the Keyfitz method with modification to incorporate the raking ratio estimation procedure.

However, for the two-stage random group design in SR areas, two alternative variance estimators given by Rao, Hartley, and Cochran (1962) and by Rao (1975) were evaluated and compared with Keyfitz's method using Monte Carlo simulation. The alternative variance estimators of estimates with and without ratio adjustment were compared with respect to their biases and stabilities. The impact on the Keyfitz variance estimator due to increase of the number of replicates was also examined. Details are reported in Section 2. Based on the results of the evaluation, the Keyfitz method was adopted for SR areas as well.

G.H. Choudhry and H. Lee, Social Survey Methods Division, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, K1A 0T6.

The Keyfitz variance formula for raking ratio estimates used for all area types in the LFS is derived in Section 3 and evaluated by Monte Carlo study. Finally in Section 4, some concluding remarks are given.

2. VARIANCE ESTIMATION FOR THE SR DESIGN

2.1 SR Design

The LFS design in the SR areas is a two-stage random group design (Rao et al. 1962) with probability proportional to size (PPS) selection of primary sampling units (PSU's) and systematic selection of dwellings at the second stage such that the design becomes self-weighting. Suppose that there are N PSU's in a given stratum and let x_j and M_j , $j = 1, 2, \dots, N$, respectively be the size measure and dwelling count for the j -th PSU in the stratum. Let $1/W$ be the sampling rate in the stratum, where W is an integer, and n be the number of PSU's to be selected from the stratum. The N PSU's in the stratum are randomly partitioned into n groups so that the i -th random group contains N_i PSU's, and $\sum_{i=1}^n N_i = N$.

Define

$$p_j = \frac{x_j}{\sum_{t=1}^N x_t}, \quad j = 1, 2, \dots, N,$$

and

$$\begin{aligned} \delta_{ij} &= 1 \text{ if the } j\text{-th PSU is in the } i\text{-th group} \\ &= 0 \text{ otherwise.} \end{aligned}$$

Then $\pi_i = \sum_{j=1}^N \delta_{ij} p_j$ is the relative size of the i -th random group.

Now define W_{ij} , the sampling interval for systematic sampling, as follows: Let $a_{ij} = \delta_{ij} W p_j / \pi_i$ and $r_{ij} = a_{ij} - [a_{ij}]$ where $[a]$ is the greatest integer less than or equal to a . Without loss of generality, we assume that the set $\{r_{ij}, j = 1, 2, \dots, N\}$ is in descending order. Then, W_{ij} is defined as

$$\begin{aligned} W_{ij} &= [a_{ij}] + 1, \quad j = 1, 2, \dots, R \\ &= [a_{ij}], \quad j = R + 1, \dots, N \end{aligned}$$

where $R = \sum_{j=1}^N r_{ij}$. Then, by definition $\sum_{j=1}^N W_{ij} = W$ for the i -th random group $i = 1, 2, \dots, n$.

Since W_{ij} is the sampling interval for systematic sampling from the selected cluster in the i -th random group, it is defined as an integer for operational simplicity.

One PSU is selected with probability proportional to W_{ij} 's from each of the n random groups independently. The selected PSU j from the i -th random group is sub-sampled systematically at the rate $1/W_{ij}$. Then the overall sampling rate in each of the n random groups is $1/W$ so that the design becomes self-weighting with a design weight equal to W . Each random group is assigned a panel number from 1 to 6. The number of random

groups n is usually a multiple of six so that each panel has the same number of random groups.

Since only one PSU is selected from each random group, we denote by $1/W_i$ the subsampling rate in the selected PSU from the i -th random group and by m_i the number of selected dwellings from the random group i .

2.2 Alternative Variance Estimators

Suppose that we are interested in the total of a characteristic y for the stratum. Let y_{jk} be the y -value for the k -th dwelling in the j -th PSU where $k = 1, 2, \dots, M_j$. Then the total $Y = \sum_{j=1}^N \sum_{k=1}^{M_j} y_{jk}$ can be estimated by $\hat{Y} = W \sum_{i=1}^n y_i$, where y_i is the sum of y -values for the m_i sampled dwellings from the PSU selected from the i -th group, $i = 1, 2, \dots, n$. We consider the following variance estimators for estimating the variance of the estimated total \hat{Y} :

(1) Keyfitz's (1957) Variance Estimator

This estimator was used in the old design with two pseudo-replicates formed by collapsing the odd numbered panels into one replicate and the even into the other. Ignoring the finite population correction (fpc), the variance is obtained by

$$\hat{V}_1(\hat{Y}) = W^2 \left(\sum_o y_i - \sum_e y_i \right)^2 \quad (2.1)$$

where \sum_o is the summation over all the odd numbered panels and \sum_e is the summation over all the even numbered panels. Alternatively, the generalized Keyfitz variance estimator for $n(\geq 2)$ replicates which is given by

$$\hat{V}_2(\hat{Y}) = W^2 \frac{n}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \quad (2.2)$$

where $\bar{y} = (1/n) \sum_{i=1}^n y_i$, can be used. In this case each PSU or panel is taken as a replicate. \hat{V}_2 was considered because it was thought that this variance estimator might have better efficiency (stability) than \hat{V}_1 due to its larger number of degrees of freedom.

(2) Rao, Hartley, and Cochran's (1962) Variance Estimator

This variance formula is derived under the assumption that the number of secondaries m_i to be selected from the i -th group is fixed for $i = 1, 2, \dots, n$, and simple random sampling (SRS) is also assumed at the second stage. The variance estimator is given by:

$$\hat{V}_3(\hat{Y}) = A \sum_1^n \pi_i \left(\frac{M_i y_i}{m_i p_i} - \hat{Y} \right)^2 + \sum_1^n \frac{\pi_i}{p_i} M_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i} \right) s_i^2 \quad (2.3)$$

where

$$A = \frac{\sum_1^n N_i^2 - N}{N^2 - \sum_1^n N_i^2}, \quad (2.4)$$

$$s_i^2 = \frac{1}{m_i - 1} \sum_{k=1}^{m_i} (y_{ik} - \bar{y}_i)^2. \quad (2.5)$$

M_i is the number of dwellings in the selected PSU from the i -th group and m_i out of M_i dwellings are selected with systematic sampling but the variance estimate is obtained under the assumption of SRS. The y -value for the k -th selected dwelling from the selected PSU in the i -th group is y_{ik} and $\bar{y}_i = y_i / m_i$.

Since $\pi_i / p_i = W / W_i$ and $M_i / m_i = W_i$, (these equalities are not strict due to the use of integer values for W_i), the variance formula (2.2) can be written as:

$$\hat{V}_3(\hat{Y}) = A \sum_1^n \pi_i \left(W \frac{y_i}{\pi_i} - \hat{Y} \right)^2 + W \sum_1^n \left(1 - \frac{m_i}{M_i} \right) M_i s_i^2. \quad (2.6)$$

(3) Rao's (1975) Variance Estimator

In this case it is assumed that m_i secondaries are selected with SRS but, since the design is self-weighting, the sample size m_i at the second stage is treated as a random variable. The variance estimator is given by:

$$\begin{aligned} \hat{V}_4(\hat{Y}) = & A \sum_1^n \pi_i \left(W \frac{y_i}{\pi_i} - \hat{Y} \right)^2 \\ & + \sum_1^n \left\{ \frac{\pi_i^2}{p_i^2} - A \left(\frac{\pi_i}{p_i^2} - \frac{\pi_i^2}{p_i^2} \right) \right\} \frac{M_i^2 s_i^2}{m_i} - \sum_1^n \frac{\pi_i}{p_i} M_i s_i^2. \end{aligned} \quad (2.7)$$

where A is defined by (2.4) and s_i^2 by (2.5). After some simplification (2.7) can be written as

$$\hat{V}_4(\hat{Y}) = \hat{V}_3(\hat{Y}) + W^2 \sum_1^n m_i s_i^2 \left\{ \left(1 - \frac{W_i}{W} \right) - A \left(\frac{1}{\pi_i} - 1 \right) \right\}. \quad (2.8)$$

We note that there is an additional term, which could be positive or negative, in the variance formula when random sample size is assumed at the second stage.

2.3 Monte Carlo Study

In order to evaluate the biases of the four variance estimators and their relative stabilities, a Monte Carlo study was carried out with 19 Labour Force strata from the Census Metropolitan Area (CMA) of Halifax using data from the 1981 census. The census data for the purpose of this study was the census sample given the long questionnaire which is 20% systematic sample of dwellings within Enumeration Areas. The sampling rate $1 / W$ was taken to be 0.04 to obtain the same expected sample size as in the actual redesigned LFS. The number of random groups within each stratum was even and was determined so that the expected sample size within random groups would be as close to 4.5 as possible to correspond to the actual LFS. The 19 strata chosen for the study are shown in Table 1 with the number of PSU's, the number of selected PSU's, the number of dwellings, and the expected sample sizes along with the corresponding totals for all the strata. Within each of the 19 strata, 1,000

Table 1
Strata Used for the Monte Carlo Study

Stratum	No. of Dwellings	No. of PSU's	No. of Selected PSU's	Expected Sample Size
1	737	49	6	29.5
2	490	33	4	19.6
3	745	45	6	29.8
4	720	34	6	28.8
5	621	37	6	24.8
6	630	38	6	25.2
7	503	31	4	20.1
8	340	23	4	13.6
9	472	33	4	18.9
10	468	33	4	18.7
11	367	28	4	14.7
12	390	23	4	15.6
13	626	36	6	25.0
14	650	39	6	26.0
15	350	22	4	14.0
16	736	46	6	29.4
17	573	35	6	22.9
18	773	48	6	30.9
19	866	64	8	34.6
total	11,057	697	100	442.3

amples were generated independently using a Monte Carlo technique, employing the random group design described in Subsection 2.1.

Let \hat{Y}_{ht} be the estimate of the total Y_h for stratum h from the t -th Monte Carlo draw, $h=1, 2, \dots, 19$, and $t=1, 2, \dots, 1,000$. Similarly \hat{V}_{jht} , $j=1, 2, 3, 4$ are the four variance estimators of \hat{Y}_{ht} .

Now define

$$Y = \sum_{h=1}^{19} Y_h,$$

$$\hat{Y}_t = \sum_{h=1}^{19} \hat{Y}_{ht},$$

$$\hat{V}_{jt} = \sum_{h=1}^{19} \hat{V}_{jht}, \quad j = 1, 2, 3, 4,$$

where $t = 1, 2, \dots, 1000$.

\hat{Y}_t is the estimate of the total Y obtained from the t -th Monte Carlo draw and \hat{V}_{jt} , $j = 1, 2, 3, 4$ are the corresponding variance estimates.

The Monte Carlo expectation and variance denoted by E^* and V^* respectively are defined for T Monte Carlo draws as follows:

$$E^*(\hat{\theta}) = \frac{1}{T} \sum_{t=1}^T \hat{\theta}_t,$$

$$V^*(\hat{\theta}) = \frac{1}{T} \sum_{t=1}^T [\hat{\theta}_t - E^*(\hat{\theta})]^2,$$

where $\hat{\theta}$ is an estimator of the unknown parameter θ and $\hat{\theta}_t$ is the estimate obtained from the t -th draw. Using these definitions, we obtain the Monte Carlo variance of the estimator \hat{Y} , $V^*(\hat{Y})$, and the Monte Carlo expectations and variances of the variance estimators \hat{V}_j , $E^*(\hat{V}_j)$ and $V^*(\hat{V}_j)$ respectively for $j = 1, 2, 3, 4$.

Now define the bias of the variance estimator \hat{V}_j by:

$$B_j = E^*(\hat{V}_j) - V^*(\hat{Y}),$$

and percent bias as:

$$PB_j = 100 \frac{B_j}{V^*(\hat{Y})}, \quad j = 1, 2, 3, 4.$$

Then the Mean Square Error (MSE) of \hat{V}_j is given by:

$$MSE_j = V^*(\hat{V}_j) + B_j^2, \quad j = 1, 2, 3, 4.$$

We define the efficiency of \hat{V}_j , relative to the Keyfitz variance estimator with two replicate (i.e., \hat{V}_1) as:

$$\text{Rel. Eff}(\hat{V}_j \text{ vs. } \hat{V}_1) = (MSE_1 / MSE_j)^{1/2}, \quad j = 2, 3, 4.$$

In this study, we consider three labour force characteristics: Employed, Unemployed, and In Labour Force. The relative biases and efficiencies of the variance estimators are reported in Tables 2A and 3A respectively for the three characteristics. We observe that, with respect to bias, the variance estimators 1 and 2 are similar and so are 3 and 4. The variance estimator 1 and 2 have very large positive biases notably for Employed and In Labour Force while 3 and 4 have relatively small biases. In efficiency comparison, the variance estimators 3 and 4 are much superior to 1 and 2 and very similar to each other. Moreover, the variance estimator 2 also performed better than 1.

The four variance estimators were also evaluated for ratio estimates by total population at the level of aggregation of all the strata. The corresponding variance estimators denoted by $\hat{V}_j^{(R)}$, $j = 1, 2, 3, 4$ were also obtained from each Monte Carlo draw by the Taylor linearization method. Then we obtained ratio adjusted version of percent biases of the four variance estimators (Table 2B) and relative efficiencies of the latter three variance estimators with respect to the first one (Table 3B).

We note that the biases of the variance estimators 1 and 2 were substantially reduced for ratio adjusted estimates especially for Employed and In Labour Force. For the variance estimators 3 and 4, the biases were also reduced for Employed and In Labour Force but there was very little change for Unemployed. Although the biases of the four variance estimators are small, the only nonsignificant bias at 5% level was that of the variance estimator 3 for In Labour Force. All the observed differences between biases were significant at 5% level except those of the variance estimators 1 and 2 for the three characteristics.

Table 2A

Percent Biases of the Variance Estimators of the Estimates of LF
Characteristic Totals without Ratio Adjustment

Characteristic	Percent Bias			
	\hat{V}_1	\hat{V}_2	\hat{V}_3	\hat{V}_4
Employed	23.4	24.5	-4.7	-6.3
Unemployed	6.3	6.6	3.7	1.2
In Labour Force	24.2	25.2	-5.1	-6.7

Table 2B

Percent Biases of the Variance Estimators of the Estimates of LF
Characteristic Totals with Ratio Adjustment

Characteristic	Percent Bias			
	$\hat{V}_1^{(R)}$	$\hat{V}_2^{(R)}$	$\hat{V}_3^{(R)}$	$\hat{V}_4^{(R)}$
Employed	3.7	4.3	-1.1	-3.1
Unemployed	5.3	5.5	4.0	1.4
In Labour Force	4.5	5.0	-0.5	-2.5

Table 3A

Relative Efficiencies of \hat{V}_2 , \hat{V}_3 , and \hat{V}_4 with Respect to \hat{V}_1
(Rel. Eff. of $\hat{V}_j = [MSE(\hat{V}_1) / MSE(\hat{V}_j)]^{1/2}$, $j = 2, 3, 4$)

Characteristic	Relative Efficiency		
	\hat{V}_2	\hat{V}_3	\hat{V}_4
Employed	1.51	3.22	3.11
Unemployed	1.52	1.71	1.76
In Labour Force	1.49	3.24	3.12

Table 3B

Relative Efficiencies of $\hat{V}_2^{(R)}$, $\hat{V}_3^{(R)}$, and $\hat{V}_4^{(R)}$ with Respect to $\hat{V}_1^{(R)}$
(Rel. Eff. of $\hat{V}_j^{(R)} = [MSE(\hat{V}_1^{(R)}) / MSE(\hat{V}_j^{(R)})]^{1/2}$, $j = 2, 3, 4$)

Characteristic	Relative Efficiency		
	$\hat{V}_2^{(R)}$	$\hat{V}_3^{(R)}$	$\hat{V}_4^{(R)}$
Employed	2.13	2.59	2.52
Unemployed	1.57	1.71	1.76
In Labour Force	2.08	2.56	2.51

Table 4
Coverage Rates of 95% Confidence Intervals for the
Estimates of LF Characteristic Totals with Ratio Adjustment

Characteristic	Coverage Rate			
	$\hat{V}_1^{(R)}$	$\hat{V}_2^{(R)}$	$\hat{V}_3^{(R)}$	$\hat{V}_4^{(R)}$
Employed	93.6	95.4	94.6	94.2
Unemployed	94.3	95.1	95.3	95.0
In Labour Force	93.2	95.3	94.6	94.2

We also computed the 95% confidence intervals (CI's) for the ratio-adjusted estimates from each Monte Carlo draw using the four variance estimators. The coverage rates were obtained as the proportion of CI's which include the true value of characteristic total. The results are given in Table 4 and show that the performances of all the 4 variance estimators are very good for all the characteristics. Since the variance estimators of ratio-adjusted estimates provide confidence intervals which have coverage rates very close to the nominal value, the small biases of the variance estimators are of no practical consequence. Thus, from the bias point of view, all four variance estimators for the ratio-adjusted estimates are not much different from each other. The relative efficiencies of the variance estimators 3 and 4 are now only marginally better than 2 regardless of characteristic. The relative efficiencies of the 3 alternatives in this case are over 2 for Employed and In Labour Force. For unemployed they are somewhat lower and lie between 1.5 and 1.8, which are almost the same as those for the unadjusted case. We should note here that the variance estimator 1 is computed with 19 degrees of freedom (1 per stratum). On the other hand, in the case of the 3 alternatives we have 81 degrees of freedom since each PSU is a replicate. Hence, we conclude that the stability of the Keyfitz variance estimator for the ratio-adjusted estimates is significantly improved by increasing the number of replicates and becomes comparable with the other two alternatives (see Table 3B).

2.4 Keyfitz's Variance Estimators with 2 vs. 6 Replicates for the LFS

The results of the Monte Carlo study reported in the previous sub-section have shown that the Keyfitz variance estimator compares well with the alternate methods for the variances of the ratio-adjusted estimates both from the bias and efficiency point of view when each method uses the same number of replicates. In addition, Keyfitz's method has the advantage of simplicity and estimating the variances of changes and averages under the alternative methods involves many complications. Therefore, the Keyfitz method was retained for the SR areas as well. In order to improve the efficiency of Keyfitz's method, 6 rotation panels were adopted as replicates as opposed to 2 replicates in the old design. One major concern with using the rotation panels as replicates was whether there would be any serious inflation of the variance estimate due to panel bias.

This aspect was investigated for the three LF characteristics by computing the variance estimates using the variance formula developed in Section 3 with 2 and 6 replicates from the actual LFS data for 24 months (March '85 - February '87). From the 24 estimated variances for each of the LF characteristics, the means and standard deviations (SD's) of the variances were obtained. The ratios of the means and SD's of the variances under the two alternatives (2 vs. 6 replicates) are averaged over 24 Census Metropolitan Areas (CMA's) and given in Table 5. The following observations can be made from the table:

Table 5
 Comparison of SR Variance Estimates with 2 vs. 6 Replicates
 per Stratum Based on CMA Data of the LFS
 Mar '85 - Feb '87

Characteristic	Average Ratio of Means of Variances (2 vs. 6)	Average Ratio of SD's of Variances (2 vs. 6)
Employed	0.997	1.813
Unemployed	0.995	1.515
In Labour Force	1.003	1.833

Note: For each CMA, means and standard deviations of variance estimates were obtained from 24 months data for 2 and 6 replicates. Then the ratios (2 rep. vs. 6 rep.) of means of variances and of standard deviations (SD's) of variances were calculated for each CMA. The average ratios in the table are the averages over 24 CMA's.

- (i) The effect on the levels of the variances due to using 6 replicates as compared to 2 is very minimal, which means that adopting rotation panels as replicates has little impact on the bias of the variance estimates.
- (ii) As expected, the variances are more stable with 6 replicates than with 2 and the results are not much different from those of the Monte Carlo study (see the first column in Table 3B)

From the above observations, we conclude that the efficiency of the Keyfitz method is improved substantially without having serious impact on the bias by adopting the 6 rotation panels as replicates as opposed to using only 2 replicates.

3. VARIANCE ESTIMATION FOR RAKING RATIO ESTIMATES

3.1 Raking Ratio Estimation for the LFS

In the old LFS, post-stratified ratio estimation was used. The subweight, which is the design weight adjusted for non-response, was ratio-adjusted to external estimates of the LFS target population for 38 post-strata defined by age and sex at provincial level. The LFS target population is the population 15 years of age and over excluding armed forces, inmates of institutions, and population living on Indian reserves.

This ratio estimation enhanced the quality of provincial data substantially but subprovincial data still had somewhat poor reliability. In order to improve subprovincial data especially for Economic Regions (ER's) and Census Metropolitan Areas (CMA's), a raking ratio estimation procedure was adopted, through which simultaneous ratio adjustment at provincial and subprovincial levels is achieved.

The raking procedure is carried out in a sequence of adjustments: first, the subweight is adjusted to the subprovincial (CMA's and Non-CMA parts of ER's) population and then the provincial level adjustment by age / sex (the number of age / sex groups were reduced from 38 to 24 in the redesigned sample) is applied to the resulting weight. This procedure is repeated once more to obtain a second pair of weights. Note that for the ER's containing CMA(s), the CMA part is excluded when defining adjustment cells for the ER's so that the subprovincial adjustment cells are mutually exclusive. Let W_0 be the subweight and let (W_1, W_2) and (W_3, W_4) be the two pairs of weights resulting from the first and second iteration respectively. Labour force characteristics are estimated using W_4 . Due to the order of adjustments, the marginal totals of W_4 at provincial age / sex groups are exactly the same as the external population estimates of the corresponding groups but the marginal totals of W_4 at

subprovincial level (ER and CMA) are not quite equal to the corresponding external population estimates. However, the differences are very small.

The special area frames, which are composed of military establishments, institutions, and remote areas, in general, do not respect the ER and CMA boundaries and hence, are treated differently during the raking procedure. Each special area type forms a stratum at the provincial level. The only exceptions are remote areas in the provinces of Quebec and Alberta where further stratification is carried out. Those ER's and CMA's which contribute to the special area frame will be called "contributing" ER's and CMA's. The special area record on the sample file are copied to each of the contributing ER's or CMA's with deflated subweights in proportion to the population of that particular type of special area in the contributing ER or CMA. The raking procedure is then carried out in the usual manner as described earlier.

3.2 Variance Formula for One-Iteration Raking Ratio Estimates

The variance formula for one-iteration raking ratio estimates is derived here. The basic methodology employed here is successive application of Taylor series approximation to the raking ratio estimates until we obtain a linear form of subweights. Then the replication formula is applied as in Woodruff (1971). The successive application of the Taylor series approximation was also used by Arora and Brackstone (1977a,b) and Brackstone and Rao (1979) to obtain variance formula of raking ratio estimates for simple random sampling of units or clusters. We have adopted this method for the stratified multi-stage PPS sampling design following Woodruff's approach.

Let $Y^{(0)}$, $Y^{(1)}$, $Y^{(2)}$ be the estimates of a labour force characteristic y in a province based on W_0 , W_1 , and W_2 , respectively. The superscripts in parentheses correspond to the subscripts of W 's.

Then $Y^{(2)}$ can be expressed as follows:

$$Y^{(2)} = \sum_a \frac{Y_a^{(1)}}{P_a^{(1)}} P_a \quad (3.1)$$

where $Y_a^{(1)} = W_1$ -weighted estimate of characteristic y for the age/sex group a in the province,

$P_a^{(1)} = W_1$ -weighted estimate of population for the age/sex group a in the province

$P_a =$ External estimate of population for the age/sex group a in the province.

Let $F_a = Y_a^{(1)} / P_a^{(1)}$. The first order Taylor approximation to F_a at $(E(Y_a^{(1)}), E(P_a^{(1)}))$ is

$$F_a \doteq \frac{E(Y_a^{(1)})}{E(P_a^{(1)})} + \frac{1}{E(P_a^{(1)})} \left\{ Y_a^{(1)} - E(Y_a^{(1)}) \right\} - \frac{E(Y_a^{(1)})}{\{E(P_a^{(1)})\}^2} \left\{ P_a^{(1)} - E(P_a^{(1)}) \right\}$$

where E denotes expectation.

Then a Taylor approximation to the variance of $Y^{(2)}$ can be written as

$$V(Y^{(2)}) = V\left(\sum_a F_a P_a\right) \doteq V\left\{\sum_a \frac{P_a}{E(P_a^{(1)})} (Y_a^{(1)} - R_{Y_a}^{(1)} P_a^{(1)})\right\} \quad (3.2)$$

where

$$R_{Y_a}^{(1)} = \frac{E(Y_a^{(1)})}{E(P_a^{(1)})}.$$

Now the W_1 -weighted estimates $Y_a^{(1)}$ and $P_a^{(1)}$ can be expressed in terms of W_0 -weighted estimates as follows:

$$\begin{aligned} Y_a^{(1)} &= \sum_s \frac{Y_{sa}^{(0)}}{P_s^{(0)}} P_s, \\ P_a^{(1)} &= \sum_s \frac{P_{sa}^{(0)}}{P_s^{(0)}} P_s, \end{aligned} \quad (3.3)$$

where s denotes a CMA or an ER or the complementary part of an ER after removing the CMA part and P_s is population of the subprovincial area s . Substituting the expressions for $Y_a^{(1)}$ and $P_a^{(1)}$ from (3.3) into (3.2) and applying the first order Taylor approximation to the ratios of W_0 -weighted estimates, we obtain

$$\begin{aligned} V(Y^{(2)}) &\doteq V \left[\sum_a \frac{P_a}{E(P_a^{(1)})} \sum_s \frac{P_s}{E(P_s^{(0)})} \left\{ \left(Y_{sa}^{(0)} - R_{Ysa}^{(0)} P_s^{(0)} \right) \right. \right. \\ &\quad \left. \left. - R_{Ya}^{(1)} \left(P_{sa}^{(0)} - R_{Psa}^{(0)} P_s^{(0)} \right) \right\} \right], \end{aligned} \quad (3.4)$$

where

$$R_{Ysa}^{(0)} = \frac{E(Y_{sa}^{(0)})}{E(P_s^{(0)})} \text{ and } R_{Psa}^{(0)} = \frac{E(P_{sa}^{(0)})}{E(P_s^{(0)})}.$$

The expression in (3.4) can be written in terms of replicate level estimates. Define

$$\begin{aligned} Z_{Yshia}^{(0)} &= \frac{P_a}{E(P_a^{(1)})} \frac{P_s}{E(P_s^{(0)})} (Y_{shia}^{(0)} - R_{Ysa}^{(0)} P_{shi}^{(0)}), \\ Z_{Pshia}^{(0)} &= \frac{P_a}{E(P_a^{(1)})} \frac{P_s}{E(P_s^{(0)})} (P_{shia}^{(0)} - R_{Psa}^{(0)} P_{shi}^{(0)}), \end{aligned} \quad (3.5)$$

where h denotes a stratum belonging to s and i denotes a replicate in h .

Then (3.4) can be rewritten by rearranging the order of summations as follows:

$$\begin{aligned} V(Y^{(2)}) &\doteq V \left\{ \sum_s \sum_{h \in s} \sum_{i=1}^{n_h} \sum_a \left(Z_{Yshia}^{(0)} - R_{Ya}^{(1)} Z_{Pshia}^{(0)} \right) \right\} \\ &= V \left(\sum_s \sum_{h \in s} \sum_{i=1}^{n_h} D_{shi}^{(0)} \right) \end{aligned} \quad (3.6)$$

where

$$D_{shi}^{(0)} = \sum_a \left(Z_{Yshia}^{(0)} - R_{Ya}^{(1)} Z_{Pshia}^{(0)} \right).$$

Apart from special area strata, $(\sum_{i=1}^{n_h} D_{pshi}^{(0)})$'s are independent because they are based on subweights. However, for the special area strata they are highly correlated because the same records are attributed to the contributing subprovincial areas.

We can rewrite (3.6) as

$$\begin{aligned} V(Y^{(2)}) &\doteq V \left(\sum_{h \in S} \sum_h \sum_{i=1}^{n_h} D_{shi}^{(0)} \right) \\ &= V \left(\sum_h \sum_{i=1}^{n_h} \sum_{s \ni h} D_{shi}^{(0)} \right) \end{aligned} \quad (3.7)$$

where $\sum_{s \ni h}$ is summation over all the subprovincial areas containing the stratum h . For a non-special stratum, the stratum appears only in one subprovincial area, and the summation $(\sum_{s \ni h})$ is redundant. However, a special area stratum could appear in several subprovincial areas and the summation $(\sum_{s \ni h})$ sums up all D -values $(D_{shi}^{(0)})$, belonging to the special area stratum.

Define

$$D_{hi}^{(0)} = \sum_{s \ni h} D_{shi}^{(0)}.$$

Then (3.7) becomes

$$V(Y^{(2)}) \doteq V \left(\sum_h \sum_{i=1}^{n_h} D_{hi}^{(0)} \right). \quad (3.8)$$

The variables, $\sum_i D_{hi}^{(0)}$, are independent since they are based on subweights. Then, ignoring the fpc, the variance can be estimated by

$$\hat{V}(Y^{(2)}) \doteq \sum_h \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (D_{hi}^{(0)} - \bar{D}_h^{(0)})^2 \quad (3.9)$$

where

$$\bar{D}_h^{(0)} = \frac{1}{n_h} \sum_{i=1}^{n_h} D_{hi}^{(0)}.$$

In this expression, however, expected values are involved and these are unknown. The variance can be approximated reasonably well by substituting expected values with their estimates and hence, from (3.9), we obtain the final form of \hat{V} as follows:

$$\hat{V} \doteq \sum_h \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (D_{hi}^{(2)} - \bar{D}_h^{(2)})^2 \quad (3.10)$$

where

$$D_{hi}^{(2)} = \sum_{s \ni h} D_{shi}^{(2)},$$

$$\bar{D}_h^{(2)} = \frac{1}{n_h} \sum_{i=1}^{n_h} D_{hi}^{(2)},$$

$$D_{shi}^{(2)} = \sum_a \left(Z_{Yshia}^{(2)} - R_{Ya}^{(2)} Z_{Pshia}^{(2)} \right),$$

$$Z_{Yshia}^{(2)} = \frac{P_a}{P_a^{(1)}} \frac{P_s}{P_s^{(0)}} \left(Y_{shia}^{(0)} - \frac{Y_{sa}^{(0)}}{P_s^{(0)}} P_{shi}^{(0)} \right)$$

$$= Y_{shia}^{(2)} - \frac{P_{shi}^{(0)}}{P_s^{(0)}} Y_{sa}^{(2)},$$

$$Z_{Pshia}^{(2)} = \frac{P_a}{P_a^{(1)}} \frac{P_s}{P_s^{(0)}} \left(P_{shia}^{(0)} - \frac{P_{sa}^{(0)}}{P_s^{(0)}} P_{shi}^{(0)} \right)$$

$$= P_{shia}^{(2)} - \frac{P_{shi}^{(0)}}{P_s^{(0)}} P_{sa}^{(2)},$$

and

$$R_{Ya}^{(2)} = \frac{Y_a^{(1)}}{P_a^{(1)}} = \frac{P_a}{P_a^{(1)}} \frac{Y_a^{(1)}}{P_a} = \frac{Y_a^{(2)}}{P_a}.$$

The formula (3.10) gives the variance for W_2 -weighted estimates of LF characteristics and requires two weights W_0 and W_2 .

3.3 Application of the One-Iteration Variance Formula to Two-Iteration Raking Ratio Estimates

The variance formula for the two-iteration raking ratio estimates can be obtained by successive application of the Taylor linearization as described in the previous section. However, the formula thus obtained is very complex. It was conjectured that the variance formula for one-iteration would be a reasonably good approximation for estimating the variance of the two-iteration raking ratio estimates. The rationale behind this conjecture was that there were only small perturbations in the weights after the first iteration. Now, the one-iteration variance formula uses the pair of weights (W_0, W_2) . However, it was decided to use (W_0, W_4) instead of (W_0, W_2) since it was found that the use of W_4 instead of W_2 does not have any impact on the CV's of LF estimates which are based on W_4 . The one-iteration variance

formula using the pair of weights (W_0 , W_4) will be referred to as the one-iteration variance estimator.

To verify our conjecture, a Monte Carlo simulation study was carried out using the 1981 Census data from the province of Nova Scotia. In each Monte Carlo sample, the LFS design was simulated through all stages of sampling and a total of 1,000 Monte Carlo samples were selected independently. For each Monte Carlo sample, the following statistics were calculated for the three labour force characteristics at subprovincial and provincial levels;

1. Two-iteration raking ratio estimate, $Y^{(4)}$.
2. Variance estimate $\hat{V}(Y^{(4)})$ using the one-iteration variance estimator and the corresponding estimate of CV.
3. 95% confidence interval (i.e., $Y^{(4)} \pm 1.96 \sqrt{\hat{V}(Y^{(4)})}$).

At the end of simulation, the average of 1,000 CV's was computed and compared with the Monte Carlo CV which is very close to the true value. The results are given in Table 6A. In all 21 cases (3 characteristics for each of 7 areas) the differences are less than 8% and in 13 cases less than 4%.

Also, the proportion of confidence intervals which cover the true characteristic value was obtained. The results are shown in Table 6B. Coverage rates for Employed and In Labour Force are very close to the nominal value in general, whereas those for Unemployed are somewhat lower but still acceptable.

It was also found that the two-iteration raking ratio estimate is nearly unbiased with a maximum of 0.35 percent bias in all 21 cases.

Table 6A
Average CV's Obtained by the
One-Iteration Variance Estimator and the Monte Carlo CV's

Characteristic	ER 210	ER 220	ER 230	ER 240	ER 250	CMA Halifax	Province (Nova Scotia)
Average CV's							
Employed	3.52	3.46	3.14	3.05	1.96	2.01	1.08
Unemployed	10.36	12.28	13.13	13.43	10.35	10.55	5.27
In Labour Force	2.98	3.17	2.85	2.73	1.77	1.83	0.91
Monte Carlo CV's							
Employed	3.48	3.35	2.95	2.86	1.97	1.99	1.11
Unemployed	10.90	12.71	13.28	13.37	11.12	11.31	5.59
In Labour Force	2.76	3.08	2.76	2.53	1.72	1.74	0.92

Table 6B
Coverage Rates of 95% Confidence Intervals
Constructed by the One-Iteration Variance Estimator

Characteristic	ER 210	ER 220	ER 230	ER 240	ER 250	CMA Halifax	Province (Nova Scotia)
Employed	94.5	92.8	94.0	94.7	94.7	94.9	92.5
Unemployed	92.1	90.7	91.4	91.8	92.7	92.7	93.1
In Labour Force	96.2	93.0	93.6	95.2	95.2	96.0	94.0

4. CONCLUSIONS

It has been shown that the Keyfitz variance estimation method for estimates without ratio adjustment (in this case it becomes just a replication method) has very large positive biases and low efficiencies while the alternatives have negligible biases and higher efficiencies for the labour force characteristics considered in this study.

However, for the ratio-adjusted estimates, all the methods considered here have negligibly small biases. It has also been shown that the efficiency of the Keyfitz method can be improved substantially and made comparable to the alternatives by increasing the number of replicates. It was demonstrated using actual LFS data that using 6 rotation panels as replicates in the Keyfitz variance estimator as opposed to 2 pseudo replicates does not introduce bias due to the phenomenon of rotation panel bias. As shown by Monte Carlo results, the one-iteration variance formula derived by the Keyfitz method using Taylor linearization gives reasonably good variance estimates for the two-iteration raking ratio estimates and has good coverage properties.

ACKNOWLEDGEMENT

The authors are grateful to the two referees, an Associate Editor and the Editor for their useful comments on the earlier version of the paper.

REFERENCES

- ARORA, H.R., and BRACKSTONE, G.J. (1977a). An investigation of the properties of raking ratio estimators: I with simple random sampling. *Survey Methodology*, 3, 62-83.
- ARORA, H.R., and BRACKSTONE, G.J. (1977b). An investigation of the properties of raking ratio estimators: II with cluster sampling. *Survey Methodology*, 3, 232-252.
- BRACKSTONE, G.J., and RAO, J.N.K. (1979). An investigation of raking ratio estimators. *Sankhyā*, Series C, 41, 97-114.
- KEYFITZ, N. (1957). Estimates of sampling variance where two units are selected from each stratum. *Journal of the American Statistical Association*, 52, 503-510.
- PLATEK, R., and SINGH, M.P. (1976). Methodology of the Canadian Labour Force Survey. Catalogue No. 71-526, Statistics Canada.
- RAO, J.N.K. (1975). Unbiased variance estimation for multi-stage designs. *Sankhyā*, Series C, 37, 133-139.
- RAO, J.N.K., HARTLEY, H.O., and COCHRAN, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society*, 24, 482-490.
- SINGH, M.P., and DREW, J.D. (1981). Redesigning continuous surveys in a changing environment. *Survey Methodology*, 7, 44-73.
- WOODRUFF, R.S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 66, 411-414.

The "AGEVEN" Record: A Tool for the Collection of Retrospective Data

PHILIPPE ANTOINE, XAVIER BRY and PAP DEMBA DIOUF¹

ABSTRACT

Because it is easy to use, the "AGEVEN" record makes it possible to date events more precisely and to classify retrospectively demographic events (births and deaths), changes in marital status and changes in place of residence. The data collected are used to accurately recreate the socio-economic conditions that were present when the demographic events being studied took place.

KEY WORDS: Retrospective survey; Biographies; Demographic survey.

1. INTRODUCTION

Two major data collection methods are available to demographers to collect data on natural movement (natality and mortality): longitudinal observation and retrospective questionnaires. The longitudinal observation method (following a population sample over a relatively long period of time) is, in theory, the method which provides the most accurate results. It does, however, have its drawbacks. It is expensive because of the amount of travel required for observation, and a relatively lengthy period of time is needed to obtain results. Finally, in urban areas, the method is difficult to apply because of the high degree of mobility of the population, which leads to a significant deterioration of the sample, such as that encountered in IFORD's infant and child mortality surveys (Scott 1985; Fargues 1985).

The retrospective method gives less reliable results because it depends more on the memory of the respondents. However, the total observation period is generally longer than that of the longitudinal surveys introduced in recent years in African countries. The risk of omitting events remains high and dating them is inaccurate. Finally, in urban areas there is a tendency when reconstituting the past to mix events which took place in the city being surveyed with other, earlier events, which took place in other places of residence (urban or rural).

Since we wished to determine mortality and fertility differences in Pikine, a suburb of Dakar, and also wished to obtain fairly reliable results quickly, we selected a data collection method that would enable us to recreate accurately the infant and child mortality risk factors at the time of death of each of the children of the women surveyed. The survey was conducted jointly by the Senegal Statistics Branch and Orstom (Antoine et Diouf 1986). The field work was carried out between March and May 1986. The first results were available in September 1986. The method we selected is different from the retrospective method most frequently used, which takes into account only the socio-economic and cultural characteristics of the women at the time of the survey. These characteristics could, in fact, have changed considerably during the women's child-bearing years (improvement or deterioration of living conditions, change of marital status, change of activity, and so forth). Our method makes it possible to better assess the relationship between urban insertion and changes in demographic behaviour. The following objectives determined our collection strategy:

to obtain a complete list of the events observed (mainly births and deaths);

¹Philippe Antoine, demographer, and Xavier Bry, statistician, ORSTOM, P.O. Box 1386, Dakar, Senegal; Pap Demba Diouf, demographer, Statistics Branch, P.O. Box 116, Dakar, Senegal.

- to date these events as accurately as possible;
- to place the events in their socio-economic context (marital status, professional status of the husband and wife, living conditions).

2. COLLECTION AND DATING OF DEMOGRAPHIC EVENTS

To conduct a successful retrospective survey means, in particular, establishing as accurate a biography as possible (in relation to the field studied) for each person surveyed. A method has to be found, therefore, to situate past events chronologically.

A number of methodological improvements have been proposed in the past. Ferry (1977) used an "event file", which involved assigning a record to each event. According to the author, the originality of this method lay in placing the events in order together with the person surveyed (pregnancies, marriages and divorces, places of residence and so forth) and situating them in relationship with each other. The technique consisted in recreating, with the person surveyed, the succession, logic, interferences and, finally, the individual biography. However, it is a relatively complex method and involves handling numerous records in the field and during processing.

Another method of classifying and dating events was used in the Senegalese survey on fertility in 1978: the "AGEVEN" graph. There were two reasons for using the "AGEVEN" graph in the Senegalese survey:

- to make it possible to better estimate the age of the women and their children with the help of relatively precise dating;
- to make it possible to accurately estimate fertility by preparing the pregnancy histories of all the women.

The "AGEVEN" graph used in the Senegalese fertility survey (Figure 1) plots two curves. The righthand curve describing the lifeline of the woman (LL curve) is graduated in intervals of three months, making it possible to plot inside a year the events affecting the woman. The lefthand curve, called the AE (age of events) curve, indicates the time which has passed between the event and the date of the survey. Thus, an age on the AE curve corresponds to each year on the LL curve, and vice versa. This graph, which was also used in the Ivory Coast fertility survey, seems to be mainly an instrument for dating events.

3. USE OF THE "AGEVEN" RECORD IN THE PIKINE SURVEY

We tried to combine some of the advantages of each of these collection methods: the "AGEVEN" graph, which is easy to use to date events, and the event file, which makes it possible to take various kinds of events and to classify them in relation to each other. We systematized the "AGEVEN" record by distinguishing between demographic events (births, deaths), changes in marital status and changes in place of residence. For convenience, we retained the name given the graph used in the Senegalese fertility survey for our record, but while the name is the same, the uses which can be made of it are different. The "AGEVEN" record (see Figure 2) contains three columns:

- the first covers demographic events (births (B); deaths (DT); abortions (A); miscarriages (MC); stillbirths (SB)). Each event (birth or death) must be followed by its chronological ranking, the first and last names of the child and, possibly, the exact date;
- the second column covers matrimonial events and the chronological ranking of each of the spouses or partners (marriages (M); divorces (D); widowhood (W), the rank of the various fathers (indicated as F1, F2, ... Fn).

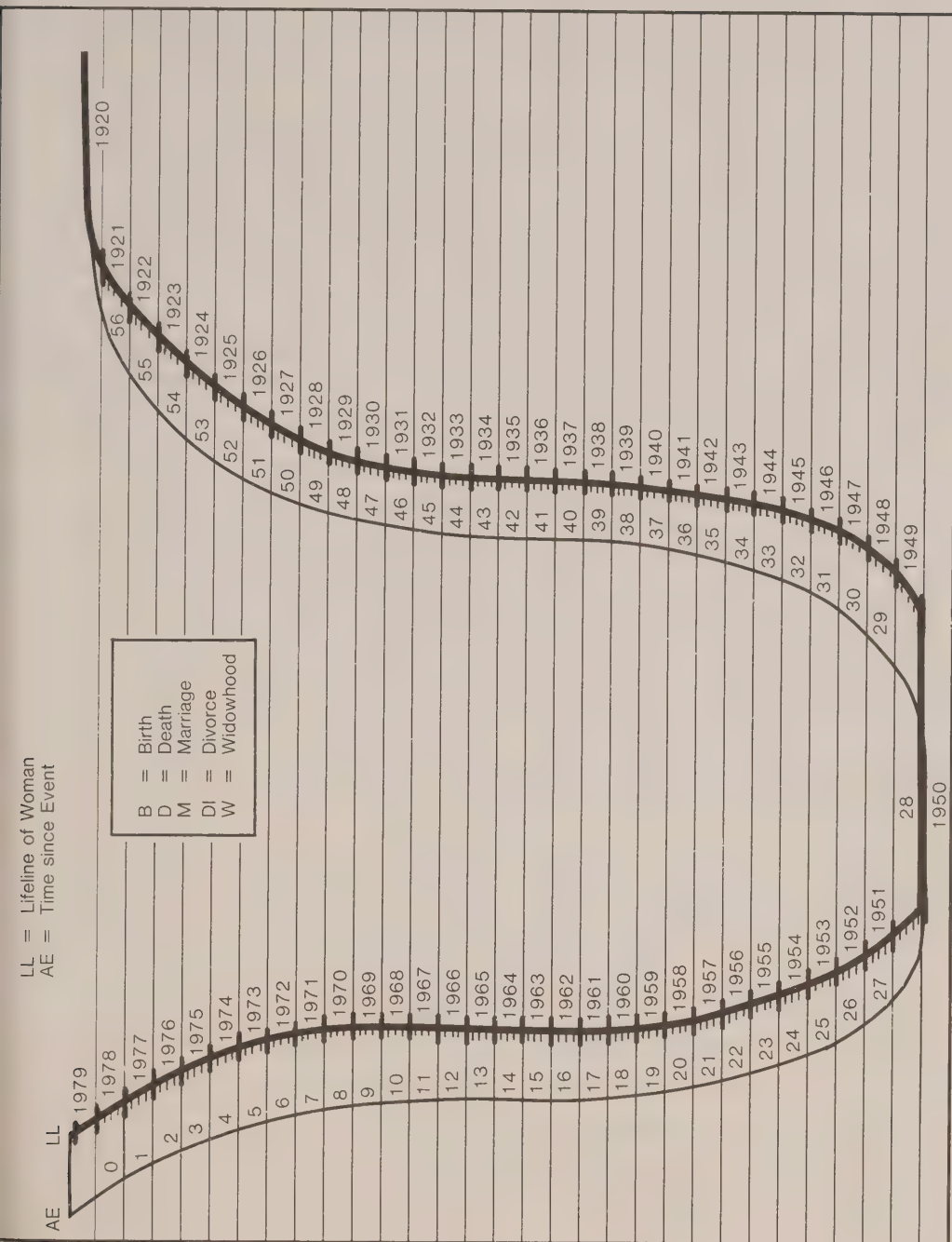


Figure 1. AGEVEN Graph Used in the Senegalese Fertility Survey

Year	Age	Demographic Events	Matrimonial Events	Places of Residence	AGEVEN	BNR/ORSTOM
1986	March				CODES B _i + date + name + place = birth child no. i D _i + date + name + place = death child no. i M _i + date + name + place = marriage no. i DI _i + date + name = divorce spouse no. i V _i + date + name = widowhood spouse no. i MC = miscarriage A = abortion SB = stillbirth F _i = father no. i	IDENTIFIER Name: Block: Concession: Household: Woman No.:
1985	0					
1984	1	B ₃ Aminata 18-12-84	F ₂	Pikine		
1983	2					
1982	3		M ₂	Pikine		
1981	4					
1980	5					
1979	6	D ₁ Ibrahima at age 4	DI ₁	Pikine		
1978	7	B ₂ Abdoul 5-01-78	F ₁	Dakar		
1977	8					
1976	9				Year Age Demographic Events Matrimonial Events Places of Residence 1936 49 1937 48 1938 47 1939 46 1940 45 1941 44 1942 43 1943 42 1944 41 1945 40 1946 39 1947 38 1948 37 1949 36 1950 35 1951 34 1952 33 1953 32 1954 31 1955 30 1956 29 BW Awa 1957 28 Kaolack	
1975	10	B ₁ Ibrahima	F ₁	Dakar		
1974	11					
1973	12		M ₁	Thiès		
1972	13					
1971	14					
1970	15					
1969	16					
1968	17					
1967	18					
1966	19					
1965	20					
1964	21					
1963	22					
1962	23			Kaolack		
1961	24					
1960	25					
1959	26					
1958	27					

Figure 2. Example of use of the "AGEVEN" record.

the third column indicates the place of residence at the time of each of these demographic and matrimonial events. This column makes it possible to follow the migratory paths of the women and to determine the date of their arrival in Pikine.

The "AGEVEN" record is a methodological tool that serves various purposes:

situating events chronologically;

helping the woman situate chronologically events for which she has forgotten the date;

ensuring that all the demographic events lived by the woman surveyed are recorded;

identifying changes of residence and the location where events took place;

checking the consistency of events among themselves.

The interview consists of two phases: one involving the household and the other involving the women between the ages of 15 and 49. The "household" questionnaire, which lists all members of the household, whether currently residing in the household or not, deals in particular with the filiation of the persons surveyed, their blood relationship with the head of the household or "nucleus," their sex, their marital status, and their date of birth or age. The "women's" questionnaire concerns all the women, resident and present in the household, between the ages of 15 and 49. The "AGEVEN" record is used to complete this questionnaire.

To transcribe the data collected on this record, the investigator can take various points of reference (the date of birth of the woman, the date of birth of her first child, and so forth) and, with the help of the respondent, reconstitute her entire lifeline, namely all the other events which have taken place during her life, such as marriage, divorce, and various pregnancies. This operation may be broken down as follows:

After recording the first live birth, the investigator asks the respondent to state all subsequent live births, in chronological order, indicating whether or not the child is still alive and whether or not he or she is still living in the household.

The investigator then records these births on the record, using the official documents shown to him. In our case, official documents were available mainly for children born in the Dakar area. For the age of the women, however, as well as for the birthdates of some children, the investigator has to rely on elements in the historical calendar to determine the dates (month and year).

The "AGEVEN" record makes it possible to situate events according to the age of the woman at the time of the event, the time which has passed since the event took place, or the date of the event. Any large gap between two births or other inconsistency between two events is easily detected during the interview with the woman.

It is also possible to use the "AGEVEN" record to check the consistency of events. For example, two children cannot be born within nine months of each other; a woman cannot say that she was married at age 12 and had her first child in 1970 at age 14, and then go on to say that she was born in 1950. In the latter case, there is likely an error in the date of birth of the woman and it should be corrected.

The record makes it possible to record both events for which an exact date is given and events for which only an age is given (such and such a child is now ten years old; I was married 15 years ago). Finally, with the help of this record, events for which the date is not clear can be situated. For example, such and such a child was born between the one born on 10-2-74 and the one born in 1978. It is highly likely that this child was born in 1976. To use this record successfully, the investigator must take a critical look at the chain of events and must try to make it as complete as possible, taking care to check the reliability and consistency of the responses provided. This is possible only if confidence is established in the dialogue with the respondent.

After having recorded all the live births declared by the respondent, the investigator turns to the intervals between successive births. All events are not always reported in the initial responses, but by using the "AGEVEN" record, the investigator can track down the

omitted events. The investigator thus asks himself what happened each time an interval of more than two years is recorded between two live births. The responses provided by the respondent may reveal abortions, stillbirths, death soon after birth, information obtained on contraceptives, and so forth. Although this was not an objective of the Pikine survey, the dialogue that is established can make it possible to delve deeper into matters relating to family planning.

Each of the events is linked to the location, marital status and partner of the woman at the time of the event. After recording all the events affecting the woman, the investigator then has to estimate more accurately the date of birth of the mother. The investigator has in fact already recorded the date of birth of the mother, as indicated either by the woman or the head of the household, when completing the "household" questionnaire. Now, in a one-on-one interview with the respondent and having recorded the events which affected her, he can provide the best possible estimate of the respondent's age.

For example. Awa was born in 1956 in Kaolack. She says that she has had three children Ibrahima, who would now be 10 years old, born in Dakar, died at age 4 in Pikine; Abdoul born on January 5, 1978 in Dakar; and Aminata, born on December 18, 1984 in Pikine. Awa was married for the first time at age 17 in Thies. She was divorced in 1979 (while living in Pikine). She remarried in 1982, at which time she was still living in Pikine (see Figure 2). During the interview, the investigator will notice a gap of almost 7 years between Abdoul and Aminata. He should ask whether there were other births or pregnancies during this period. In the case of Awa, the divorce and subsequent remarriage three years later may explain the gap. However, the investigator must check with the woman to ensure that the gap does not hide other demographic events.

The interactive form of the interview seems to encourage dialogue with the respondent and improves contact between the investigator and respondent, which is unfortunately too often clouded by doubt on the part of the investigator and mistrust on the part of the respondent Bonnet (1984). As the investigator continues his or her investigation, new events are mentioned. When he or she asks whether there was another event between two births separated by more than two years, the respondent is often surprised and responds in one of two ways. If no event has occurred, she asks, "Why do you ask that?" If, however, an event has indeed occurred, she often asks, "Who told you that?" since she has the impression that the investigator already knows something. The "AGEVEN" record becomes a kind of crystal ball, like the cowry shell. Sometimes the interview becomes a game, and the respondent is pleased to place past events in order. A woman with a complicated marital and reproductive history may even want a copy of her "AGEVEN" record. As in any survey there are problems with the use of this record. Sometimes it is difficult or awkward to be alone with the respondent, and often women are embarrassed if the record brings up events concerning a partner preceding the current husband.

In practice, the record is incomplete because there is no question which eliminates possible confusion between stillbirths and infants who die shortly after birth. This kind of confusion often arises in responses given in the Wolof language, in which it is difficult to distinguish between miscarriages and abortions and between stillbirths and deaths immediately after birth. Some French terms or words cannot be translated directly into Wolof. For stillbirths, for example, there is no single question that elicits the desired response. At least two questions are therefore required. When confronted with an interval between successive births, the investigator asks the following question, for example: "Lou am dikhane té Moussa ak Ali?" (what happened between Moussa and Ali?). This question correctly leads the women to stillbirths, abortions, miscarriages and so forth. To elicit a satisfactory response, clarifications are needed: "Dikhane té Moussa ak Ali, amo fi dom diou dé guinaw bou mou in bakhane?" (did you have a child who died after giving some sign of life between Moussa and Ali?). The confusion results mainly from the fact that the distinction between a miscarriage and stillbirth is not always clear and from the fact that a child is not given a name

until he or she is a week old. Also, for certain ethnic groups, it is not until the child has a name that he or she is really taken into account. A column indicating whether or not the infant cried at birth would therefore have been very useful.

The "AGEVEN" record used in the Pikine survey did indeed provide more satisfactory data than the graph used in the Senegalese fertility survey, in terms of both the nature and quantity of data collected. However, it did not eliminate the tendency to round off the intervals between successive births in years (approximately 37% of the intervals), particularly in intervals of two years, which account for approximately 20% of the intervals observed between successive births. In addition, it was not possible, using this technique, to list all the issue of young girls who had been pregnant but who had had no live births. Some biases, which are certainly classic in demography, do persist therefore, and this method does not eliminate the need to take extreme care in the field.

4. TRANSCRIPTION FROM THE "AGEVEN" RECORD TO THE QUESTIONNAIRE AND ELECTRONIC DATA PROCESSING

The questionnaire regarding the reproductive history of the women was designed in such a way as to permit the best possible transcription of the data collected using the "AGEVEN" record. First, the characteristics of each of the children are noted in chronological order by birth, along with the date of death, if appropriate. The investigator then records the marital status at the time of each of these events in order to note any possible change in spouse. Then, changes in the socio-economic situation of the father and mother are taken into account, as well as changes in living conditions and in place of residence. The survey also included other questionnaires regarding the characteristics of the household, individuals and women observed.

The data collection method allows for two kinds of analysis. The first involves a classical analysis of mortality by generation and sub-population (according to neighbourhood, type of housing and so forth). However, what is especially interesting about this study is that it allows for analysis of mortality (and fertility) taking into account migratory behaviour and changes in the socio-economic conditions of the women surveyed. When this method is used, mortality is no longer interpreted solely according to the socio-economic conditions at the time of the survey. Rather, it is related to the conditions which really existed at the time of the event, and it is therefore possible to better understand the differences relating specifically to living conditions in urban areas (Pikine in this case).

Depending on the place of birth of the child, different mortality rates were recorded. Many of the respondents are migrant women from other cities or from villages in the interior of the country. Children born to them in rural areas suffered a significantly higher risk of mortality than those born in the Dakar area.

The child mortality rate (between 1 and 4 years) clearly reveals the risks resulting from socio-economic differences. The risk of dying between the ages of 1 and 4 is 2.84 times higher for children born in villages than for those born in Pikine. The z -test shows that the difference between the two rates (Pikine mortality rate and rural mortality rate) is significant. We tested the hypothesis that the mortality rate for children born in Pikine is the same as that for children born in rural areas. Since the sample sizes are relatively large, approximation using the normal distribution is justified. Under the hypothesis that the mortality rates

Table 1
Mortality by place of birth (in thousands)

	Pikine	Dakar	Other Cities	Rural	Total	Pkn-Rural Test
Infants	52	57	45	114	58	-6,586**
Children	55	62	90	156	68	-10,093**
Population	5155	1513	644	704	8016	

are equal, the z -statistic is distributed as a standard normal variable. The symbol “***” indicates a significant difference at the $\alpha = 0.05$ level. Classic retrospective data collection without distinction as to the place of birth of the child would have led us to class births outside Pikine with those inside Pikine and would have resulted in a higher mortality rate (child mortality rate of 68 per thousand rather than 55 per thousand).

Moreover, a second analysis can be made for each of the women observed. A simplified biographical file can be created in which the successive stages are defined in terms of births. A relationship is thus established between matrimonial events, changes in residence and reproductive data. The principal stages in the migratory path followed since the birth of the first child, or since marriage, can also be reconstructed. Longitudinal data gathered in this way lend themselves very well to recent methods for the analysis of interference between phenomena (Courgeau and Lelievre (1986); Cox and Oakes (1984)).

5. CONCLUSION

The data collected for each of the variables are very brief, but they should make it possible to detect some significant differences and to determine the living conditions at the time of birth and death. The collection methodology used is adapted to the collection of data on the reproductive histories of the women and the destiny of their children. The main advantage of the “AGEVEN” record is its facility in pinpointing various events chronologically and in classifying these events in relationship with each other, without eliminating the possibility of inserting events omitted as the interview proceeds. The flexibility of the “AGEVEN” record leads us to suggest that it could be used in other fields, for professional biographies or migratory routes, for example, by establishing a parallel between place of residence, profession, marital status, family situation, living conditions and so forth. A great deal of methodological research has been conducted in the analysis of demographic biographies (Courgeau 1984; Haeringer 1972; Riandey 1985). Our method is intended merely as a simple and reliable tool for the collection of data. It is up to each user to determine which variable he or she wishes to arrange chronologically using the “AGEVEN” record and, once the biographical framework has been collected, to obtain more data on the field(s) he or she is studying, using the questionnaire.

ACKNOWLEDGMENTS

The authors would like to thank the referees for their helpful comments.

REFERENCES

- ANTOINE, Ph., and DIOUF, P.D. (1986). Changements démographiques en milieu urbain. Paper presented at Séminaire sur la mortalité au Sénégal. Dakar.
- BONNET, D. (1984). Occultation, omissions. Quelques problèmes soulevés par l'enquête quantitative en matière de santé. *Medicus Mundi*, 11.
- COURGEAU D. (1984). Relations entre cycle de vie et migrations. *Population*, 39, 483-513.
- COURGEAU, D., and LELIEVRE, E. (1986). Nuptialité et agriculture. *Population*, 41, 303-326.
- COX, R., and OAKES, D. (1984). *Analysis of Survival Data*. London: Chapman and Hall.
- DIRECTION DE LA STATISTIQUE (1981). *Enquête Sénégalaise sur la Fécondité, 1978 - Rapport National d'Analyse*, 1.
- FARGUES, Ph. (1985). L'évaluation du niveau de la mortalité à partir des données des enquêtes EMIJ. *Les enquêtes sur la mortalité infantile et juvénile (EMIJ)*, 1, 60-84.
- FERRY, B. (1977). Le fichier événement. Une nouvelle méthode d'observation rétrospective. In *l'Observation démographique dans les pays à statistiques déficientes*. Liege, Belgium: Ordina Editions, 137-150.
- HAERINGER, Ph. (1972). Méthodes de recherche sur les migrations africaines. Un modèle d'interview biographique et sa transcription synoptique. *Cahiers ORSTOM*, 9, 439-453.
- SCOTT, Ch. (1985). Les problèmes de déperdition dans les enquêtes suivies. In *Les enquêtes sur la mortalité infantile et juvénile (EMIJ)*, 1, 44-47.
- RIANDEY, B. (1985). L'enquête "biographie familiale professionnelle et migratoire" (INED, 1981). Le bilan de la collecte. In *Migrations internes, collecte des données et méthode d'analyse. Département de démographie*. Université de Louvain, 117-134.

An Alternative Method of Controlling Current Population Survey Estimates to Population Counts

K.R. COPELAND, F.K. PEITZMEIER, and C.E. HOY¹

ABSTRACT

The CPS uses raking ratio estimation in post-stratification estimation to adjust sample estimates of population to census-based estimates of the population. An alternative procedure, using generalized least squares, is compared to the current procedure.

KEY WORDS: Generalized least squares; Post-stratification; Raking ratio estimation.

1. INTRODUCTION

The Current Population Survey (CPS) produces labor force estimates for the total U.S. working-age civilian noninstitutional population, based on a monthly multi-stage probability sample of approximately 60,000 housing units in the U.S. Each month a rotating sample comprised of 8 panels (called rotation groups) of housing units is interviewed, with demographic and labor force data being collected for all civilian adult occupants of the sample housing units.

Monthly estimates are published, subaggregated by demographic characteristics. Estimates for other subaggregates of the population (states, families, veterans, wage and salary earners, persons not in the labor force, etc.) are also produced on a monthly, quarterly, and/or annual basis.

Sample person weights are derived through the application of probability of selection, adjustment for nonresponse, and ratio adjustment to reduce the contribution to the variance due to the sampling of primary sampling units. A post-stratification estimation procedure adjusts the sample person weights so as to control the survey estimates of population to independently derived estimates of the population. The resultant weights are used in a composite estimation procedure and then seasonally adjusted to produce national estimates (Hanson 1978).

Detailed estimates for certain population subdomains (families, wage and salary earners, persons not in the labor force, family earnings, and veterans) make use of sample weights derived from adjustment procedures built on top of the post-stratification estimation.

The use of a generalized least squares (GLS) approach could potentially be used in place of post-stratification estimation or to integrate the various CPS adjustment procedures. The use of GLS has been proposed and investigated for use in the Consumer Expenditure Survey (Zieschang 1986).

This article discusses and compares the current CPS post-stratification estimation (which uses raking ratio estimation) and the GLS procedure, based on two months' CPS data (July 1983 and July 1984). Both macro and micro level data were examined to evaluate differences, if any, in the two procedures in this application.

¹ K.R. Copeland, F.K. Peitzmeier, and C.E. Hoy, Division of Statistical Methods, Office of Employment and Unemployment Statistics, Bureau of Labor Statistics, Washington, D.C. 20212 U.S.A.

2. CURRENT CPS POST-STRATIFICATION ESTIMATION

The CPS post-stratification estimation uses raking ratio estimation (RRE) to adjust the sample weights within a rotation group so as to control the sample estimates for the population to independently derived estimates of the population in each of three categories (state/age/sex/ethnicity, age/sex/race).

The methodology for RRE was first proposed by Deming and Stephan (1940) as an iterative alternative to least squares adjustment of table data. The RRE procedure has been shown to produce best asymptotically normal (BAN) estimates under simple random sampling, and to minimize the adjustments made to the sample weights based on one measure of closeness, as discussed in subsection 4.2 (Ireland and Kullback 1968). In addition, RRE, although producing biased estimates, can sometimes be effective in reducing the mean square error of survey estimates. This is believed to be the case in the application of RRE for CPS (Hanson 1978).

For the CPS, the RRE procedure attempts to adjust the sample counts $\{n_{ijk}\}$ obtained from previous stages of weighting to adjusted sample counts $\{\tilde{n}_{ijk}\}$ under the condition that

$$(A) \quad \sum_{j,k} \tilde{n}_{ijk} = m_{i..}$$

$$(B) \quad \sum_{i,k} \tilde{n}_{ijk} = m_{.j.}$$

$$(C) \quad \sum_{i,j} \tilde{n}_{ijk} = m_{..k}$$

be satisfied simultaneously,

where i = state ($i = 1, \dots, 51$),
 j = age/sex/ethnicity ($j = 1, \dots, 16$),
 k = age/sex/race ($k = 1, \dots, 70$),
 $m_{i..}$ = independent state estimate,
 $m_{.j.}$ = independent age/sex/ethnicity estimate,
 $m_{..k}$ = independent age/sex/race estimate.

The RRE procedure proportionately ratio adjusts the sample data each way (i.e., state/age/sex/ethnicity, and age/sex/race) of the table in successive steps, as follows.

(1) Ratio adjustment by state:

$$n_{ijk}^{(1,1)} = (m_{i..}/n_{i..}) n_{ijk} = a_i^{(1)} n_{ijk}.$$

(2) Ratio adjustment by age/sex/ethnicity:

$$\begin{aligned} n_{ijk}^{(1,2)} &= (m_{.j.}/n_{.j.}^{(1,1)}) n_{ijk}^{(1,1)} = b_j^{(1)} n_{ijk}^{(1,1)} \\ &= a_i^{(1)} b_j^{(1)} n_{ijk}. \end{aligned}$$

(3) Ratio adjustment by age/sex/race:

$$\begin{aligned} n_{ijk}^{(1,3)} &= (m_{..k}/n_{..k}^{(1,2)}) n_{ijk}^{(1,2)} = d_k^{(1)} n_{ijk}^{(1,2)} \\ &= a_i^{(1)} b_j^{(1)} d_k^{(1)} n_{ijk}, \end{aligned}$$

where $n_{i..}$ = sample row total
 $n_{.j.}$ = sample column total
 $n_{..k}$ = sample layer total.

The completion of the three adjustment steps constitutes one iteration of the raking process. The three steps are repeated substituting the current value of $n_{ijk}^{(h,3)}$ (adjusted sample count following the third way rake of the h -th iteration) for n_{ijk} in step (1) each time until 6 iterations are completed. (The number of iterations used in CPS was determined based on the convergence properties of the RRE for CPS and the relative gains achieved by number of iterations.) The final $\{n_{ijk}^{(6,3)}\}$ is taken as $\{\tilde{n}_{ijk}\}$.

In order to adjust the sample weights, the adjustment factor for sample records in cell $\{ijk\}$ is

$$F_{ijk} = n_{ijk}^{(6,3)} / n_{ijk}$$

$$= \prod_{h=1}^6 a_i^{(h)} b_j^{(h)} d_k^{(h)}.$$

The sample weights prior to RRE are multiplied by the appropriate F_{ijk} to obtain the adjusted weights.

3. APPLICATION OF THE GLS IN THE CPS

The generalized least squares (GLS) procedure adjusts the sample weights from prior stages of weighting by minimizing the weighted squared adjustments, subject to a set of linear 'control' constraints the adjusted weights must satisfy. This is the problem which Deming and Stephan attempted to address in developing the RRE. The GLS procedure, like RRE, produces BAN estimates under certain conditions, in this case when all the cells are nonempty (Neyman 1949). GLS, by definition, minimizes the adjustments to the sample weights based on one measure of closeness (see subsection 4.2).

For the CPS, each dimension that defines a set of controls in the current post-stratification will define a set of linear constraints for the GLS procedure. The function to be minimized is

$$f(\underline{F}) = (\underline{F} - \underline{P})' P_0^{-1} (\underline{F} - \underline{P})$$

$$= \sum_i (W_{2i} - W_{1i})^2 / W_{1i},$$

subject to $X'F = N$,

where \underline{F} = $(n \times 1)$ vector of derived final weights (W_{2i}) for each of the n sample persons,

\underline{P} = $(n \times 1)$ vector of sample person weights prior to post-stratification (W_{1i}),

P_0 = $(n \times n)$ diagonal matrix with the W_{1i} on the diagonal,

X = $(n \times k)$ design matrix whose rows correspond to sample persons, and whose columns correspond to control cells. The entries of the matrix (x_{ij}) are 0's or 1's, indicating the appropriate control categories for each of the n sample persons.

\underline{N} = $(k \times 1)$ vector of independent population estimates, corresponding to the columns of X . These estimates are the same as those used in the CPS RRE.

The columns of X are required to be linearly independent so that an inverse of the matrix $(X' P_0 X)$ is achievable. In setting up matrices X and \underline{N} for CPS, the 137 control cells used in the RRE (state, age/sex/ethnicity, age/sex/race) were reduced to a set of $k = 132$ linearly independent cells.

The unique solution to $X' \underline{F} = \underline{N}$ that minimizes $f(\underline{F})$ is, as shown in Luery (1986)

$$\underline{F} = \underline{P} + P_0 X (X' P_0 X)^{-1} (\underline{N} - X' \underline{P})$$

Although the elements of \underline{F} are not constrained to be positive, in this application of GLS for CPS, the elements of \underline{F} were all positive without the need for additional constraints. Methodology for providing non-negative weights in this context is discussed in Huang and Fuller (1978) and Zieschang (1986), among others.

4. RESULTS

4.1 Macro-Level

a. Estimates

Labor force estimates were tabulated for several demographic groups for July 1983 and July 1984, using the final weights derived from RRE and GLS. Standard errors for both RRE and GLS were calculated using a random group estimator of the form Wolter (1985)

$$\sum_{k=1}^8 (8Y_k - \hat{Y})^2 / 56,$$

where Y_k = sum of the weights for sample records from the k -th rotation group with the characteristic Y ,

\hat{Y} = sum of the Y_k .

This variance estimator, while not accounting for the multi-stage design of the CPS, was used due to the unavailability of design information on the CPS public use microdata file.

Relative differences were calculated for both estimates of level and estimates of standard error. The relative difference was defined as:

$$(Y_{GLS} - Y_{RRE}) / Y_{RRE},$$

where Y_{RRE} = estimate of Y based on the weights derived through the use of RRE,
 Y_{GLS} = estimate of Y based on the weights derived through the use of GLS.

As the data in Table 1 indicate, neither weighted labor force estimates nor estimates of standard error based on the current CPS RRE procedure and the GLS procedure showed any noticeable differences or trends when subaggregated to the sex by race/ethnicity level.

For labor force estimates by sex by race/ethnicity the estimated absolute relative difference between the CPS RRE and GLS estimates were all less than 0.3% (well below the estimated CVs of each estimate). For the majority of these estimates, in particular for total and whites the absolute relative difference was less than 0.1%.

For many of the characteristics the sign of the relative difference changed from 1983 to 1984; thus there does not appear to be a pattern to the differences in the estimates obtained from the two procedures.

Table 1
Labor Force Estimates by Sex/Race or Ethnicity

		1983				1984			
		GLS		(GLS-RRE)/ RRE		GLS		(GLS-RRE)/ RRE	
		Total (000)	S.E. (000)	Total (%)	S.E. (%)	Total (000)	S.E. (000)	Total (%)	S.E. (%)
Total									
Total	Emp	103516	403	0.00	-0.14	107535	352	-0.01	1.12
	UE	10669	221	-0.04	-0.75	8765	118	-0.06	-0.21
	Rate	9.34%	0.19%	-0.04	-0.56	7.54%	0.09%	-0.05	0.27
	NILF	59938	373	0.01	-0.68	60080	419	0.02	0.41
White									
White	Emp	91338	344	0.00	-0.33	94417	274	0.00	0.70
	UE	7928	236	0.00	-0.27	6282	120	0.00	-0.14
	Rate	7.99%	0.23%	0.00	-0.26	6.24%	0.10%	0.00	-0.16
	NILF	51915	340	0.00	-0.36	51700	358	0.00	0.39
Black									
Black	Emp	9871	69	0.06	-3.44	10371	98	0.02	0.17
	UE	2434	68	-0.12	-1.07	2202	60	-0.03	1.41
	Rate	19.78%	0.55%	-0.14	-1.60	17.51%	0.42%	-0.04	1.49
	NILF	6628	26	-0.04	-1.47	6765	109	-0.02	0.09
Hispanic									
Hispanic	Emp	6132	73	-0.03	-0.59	6607	102	-0.03	1.90
	UE	920	79	-0.05	-0.29	786	70	-0.08	-0.03
	Rate	13.04%	1.10%	-0.02	-0.33	10.63%	0.96%	-0.05	0.35
	NILF	3760	31	0.05	-0.39	3786	73	0.04	1.02
Male									
Male									
Total	Emp	58985	147	0.00	-1.58	61045	188	0.00	1.74
	UE	5980	134	-0.05	-0.88	4682	79	-0.02	0.77
	Rate	9.20%	0.19%	-0.05	-0.79	7.12%	0.11%	-0.02	1.30
	NILF	17495	178	0.01	-1.81	17840	214	0.02	0.64
White									
White	Emp	52674	482	0.00	0.42	54261	111	0.00	0.34
	UE	4484	131	0.01	-0.49	3394	93	0.01	-0.12
	Rate	7.84%	0.21%	0.00	-0.47	5.89%	0.15%	0.01	-0.13
	NILF	14985	160	-0.02	-0.40	15077	150	0.00	0.16
Black									
Black	Emp	5047	56	0.07	-1.70	5263	84	0.01	-0.50
	UE	1300	45	-0.20	-1.87	1137	33	0.08	1.12
	Rate	20.49%	0.71%	-0.21	-2.02	17.76%	0.51%	0.05	0.94
	NILF	2097	40	-0.04	-0.13	2236	88	-0.07	-0.48
Hispanic									
Hispanic	Emp	3781	48	0.01	-0.86	4064	79	-0.02	1.29
	UE	534	45	-0.16	-0.83	451	41	-0.05	0.51
	Rate	12.38%	0.99%	-0.15	-0.89	9.99%	0.95%	-0.03	0.66
	NILF	981	42	0.00	-0.42	964	57	0.07	1.40
Female									
Female									
Total	Emp	44531	320	-0.01	-0.01	46490	194	-0.01	1.48
	UE	4689	107	-0.04	-0.19	4083	88	-0.10	-1.22
	Rate	9.53%	0.23%	-0.03	-0.02	8.07%	0.16%	-0.09	-0.80
	NILF	42443	287	0.01	-0.26	42240	217	0.02	0.34
White									
White	Emp	38664	315	0.00	-0.29	40156	191	0.00	0.66
	UE	3444	115	-0.01	0.16	2888	68	0.00	-0.32
	Rate	8.18%	0.28%	-0.01	0.11	6.71%	0.15%	0.00	-0.34
	NILF	36929	283	0.01	-0.32	36623	214	0.00	0.53
Black									
Black	Emp	4824	57	0.05	0.56	5108	50	0.02	1.69
	UE	1134	46	-0.02	0.07	1065	46	-0.14	-0.62
	Rate	19.03%	0.80%	-0.06	0.08	17.25%	0.67%	-0.13	-0.63
	NILF	4531	24	-0.04	2.99	4529	59	0.01	1.49
Hispanic									
Hispanic	Emp	2350	44	-0.08	-0.46	2543	38	-0.05	3.04
	UE	385	41	0.10	0.51	335	34	-0.13	-0.62
	Rate	14.08%	1.46%	0.16	0.57	11.64%	1.18%	-0.07	-0.11
	NILF	2778	33	0.07	-0.87	2822	27	0.03	0.13

The absolute relative differences between the CPS RRE and GLS estimates of standard errors for national labor force estimates were all less than: 1.9% for total population; 0.7% for whites; 3.5% for blacks; and 3.1% for Hispanics.

b. Month-in-Sample Indexes

It is a well-documented fact that the estimates produced from the CPS final weights have certain patterns of relative bias based upon the time the rotation group has been in sample (Bailar 1975). Month-in-sample indexes

$$I_k = (8Y_k / \hat{Y}) \times 100,$$

were calculated for both July 1983 and July 1984 based upon both the RRE estimates and the GLS estimates.

Month-in-sample indexes for labor force by race, labor force by sex, and labor force by ethnicity were virtually identical for estimates based upon the CPS RRE and GLS procedures.

4.2 Micro-Level

a. Adjustments to Sample Weights

Both RRE and GLS minimize some measure of closeness between the pre- and post-adjustment sample weights. For RRE the measure is (Ireland and Kullback 1968)

$$M_A = \sum_i W_{2i} \ln (W_{2i} / W_{1i}).$$

For GLS, the measure is (Luery 1986)

$$M_B = \sum_i (W_{2i} - W_{1i})^2 / W_{1i},$$

where W_{1i} = weight for sample record i prior to adjustment,
 W_{2i} = weight for sample record i following adjustment.

Tabulation of the measures of closeness (summarized in Table 2) provided some interesting and, in some cases, puzzling results. The CPS RRE yielded smaller values for both measures. The GLS procedure did tend to produce smaller values for the measures for certain subgroup most notably for blacks and Hispanics. It should be noted that the differences between the values for the measures for RRE and GLS were almost always less than 1%.

Although M_B should be minimized through the use of the GLS procedure, the value of M_B based upon the GLS weights for the total sample was greater than the value of M_B for the CPS RRE weights for 11 of the 16 rotation groups.

In seeking a reason for this apparent contradiction, it was noted that the CPS RRE has yet to converge to the age/sex/ethnicity controls after six iterations. The extent of this non-convergence is *very small*; less than 1.0% for all control categories. However, given the difference in M_B between the RRE and GLS, a change in the RRE sample weights of on 0.1%-0.2% could reverse the results. Rerunning RRE using 15 iterations, although still not achieving convergence did provide indications that the slight lack of convergence of the RRE is the reason for the results for M_B . (It should be noted that the GLS procedure minimizes M_B among the class of adjustment procedures yielding estimates that meet the population controls. Since the CPS RRE did not converge to the population controls, it is not a member of this class.)

Table 2
Comparison of measures of closeness
based on 8 RGs for each year
(# of RGs with RRE < GLS)

	M_A		M_B	
	1983	1984	1983	1984
Total	8	8	4	7
White	7	7	3	4
Black	3	3	1	1
Hispanic	0	0	0	0
Male	2	7	1	5
Female	8	8	8	8

Although an adjustment procedure such as RRE or GLS may minimize some measure of closeness for the total sample, it does not necessarily minimize that measure of closeness for subaggregates of the sample which were controlled for (e.g., blacks, Hispanics, males). Given the use of controls, and the fact that the overall measure of closeness is being minimized, it would seem desirable to have an adjustment procedure produce small measures of closeness at the subaggregate level also. The GLS procedure yielded smaller measures in almost every rotation group for Hispanics, in many rotation groups for blacks, and in several rotation groups for whites and males.

b. Comparison of Adjustments

Both RRE and GLS determine adjustment factors within cells defined by the intersection of the marginal constraints. Each sample record within a cell receives the same factor. To compare the adjustments made by the two procedures, the factors determined for each sample record by each procedure were compared using the following ratio

$$RRE/GLS = [(W_{2i}/W_{1i})_{RRE}] / [(W_{2i}/W_{1i})_{GLS}].$$

This ratio indicates the relationship between the adjustments made to a sample person weight by the RRE and GLS procedures. For comparison purposes, values of RRE/GLS less than 0.95 or greater than 1.05 were used to denote differences in the adjustments made by RRE and GLS.

For each set of independent population controls, ratios E/C (i.e., coverage rates), where E is the sample estimate based on the sample person weights prior to post-stratification and C is the independent control, were derived.

Within each set of controls (state, age/sex/ethnicity, age/sex/race) sample records were categorized by their coverage rates. Table 3 provides the sample distribution by coverage rate categories and by the RRE/GLS values, as well as the proportion of records within each coverage rate category that have the RRE/GLS values.

The data in Table 3 indicate that, for each set of controls, sample records from population groups which were over- or under-covered to some extent by the survey (i.e., for which the coverage rate is not near 1) were more likely to be adjusted differently by RRE and GLS than were sample records in population groups adequately covered by the survey.

Table 3
Comparison of RRE and GLS adjustments, 1984

Control Marginal	Coverage Rate Category	Proportion of Total Sample	Proportion of Sample with RRE/GLS <0.95 or >1.05	Proportion of Category with RRE/GLS <0.95 or >1.05
Age/Sex/ Race	<0.7	0.007	0.057	0.219
	0.7-0.8	0.022	0.116	0.136
	0.8-0.9	0.241	0.147	0.019
	0.9-1.1	0.699	0.504	0.019
	1.1-1.2	0.021	0.069	0.084
	>1.2	0.010	0.106	0.275
Age/Sex/ Ethnicity	<0.7	0.010	0.078	0.198
	0.7-0.8	0.014	0.032	0.058
	0.8-0.9	0.106	0.135	0.033
	0.9-1.1	0.869	0.741	0.022
	1.1-1.2	0.001	0.007	0.202
	>1.2	0.001	0.007	0.373
State	<0.7	0.056	0.068	0.031
	0.7-0.8	0.111	0.180	0.042
	0.8-0.9	0.278	0.325	0.030
	0.9-1.1	0.479	0.342	0.018
	1.1-1.2	0.026	0.009	0.009
	<1.2	0.049	0.077	0.040

4.3 Computer Resources

The CPS RRE and GLS procedures were run on an IBM System 370 at the National Institutes of Health using PROC MATRIX in the SAS System. The CPU time to prepare the files and perform the weighting was approximately three times as much for the GLS procedure than it was for the RRE procedure. There was also more storage of files involved with the GLS procedure. (The size of the matrices involved for CPS are quite large, with the number of rows for \underline{P} , \underline{P}_0 , \underline{X} , and \underline{N} being around 14,000 for each rotation group.

5. SUMMARY AND CONCLUSIONS

This investigation was intended to provide a comparison of RRE and GLS as applied to the CPS, at both the macro and micro level.

The results obtained at the macro level do not indicate any difference in the estimate obtained from the RRE and GLS procedures.

The measures of closeness indicated that the CPS RRE made slightly smaller changes overall to the sample weights to meet the control constraints than did the GLS. The CPS RRE tended to produce slightly larger measures of closeness for subaggregates of minority populations. The two procedures differ most notably in the adjustments made to portions of the population which are either over- or under-covered.

Based on the work done in this investigation, it does appear that the RRE takes less computer time to run for the CPS second-stage adjustment than the GLS.

ACKNOWLEDGEMENT

The authors are grateful to Fritz Scheuren for his review of the original version of this paper, and to the referees and the Associate Editor for their very useful comments, incorporation of which resulted in the improvement of the paper.

REFERENCES

- BAILLAR, B. (1975). The Effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23-30.
- DEMING, W.E., and STEPHAN, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11, 427-444.
- HANSON, R.H. (1978). The Current Population Survey design and methodology. Technical Paper 40, U.S. Bureau of the Census.
- HUANG, E.T., and FULLER, W.A. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of the Section on Social Statistics, American Statistical Association*, 300-305.
- RELAND, C.T., and KULLBACK, S. (1968). Contingency tables with given marginals. *Biometrika*, 55, 179-188.
- QUERRY, D. (1986). Weighting sample survey data under linear constraints on the weights. *Proceedings of the Section on Social Statistics, American Statistical Association*, 325-350.
- NEYMAN, J. (1949). Contribution to the Theory of the X^2 Test. In *Proceedings of the First Berkeley Symposium on Mathematical Statistics and Probability*, (Ed. J. Neyman), Berkeley: University of California Press, 239-273.
- WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.
- CHIESCHANG, K.D. (1986). A Generalized least squares weighting system for the Consumer Expenditure Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 64-71.

A Class of Methods for Using Person Controls in Household Weighting

CHARLES H. ALEXANDER¹

ABSTRACT

A class of "constrained minimum distance" methods is considered for constraining household weights to be consistent with auxiliary information on the number of persons in various age \times race \times sex cells. The constrained weights are as close as possible to the initial weights based on the inverse probability of selection. This class of methods includes raking and generalized least square methods, as well as multinomial maximum likelihood, (where the cells of the distribution are household types.) The properties of the methods in the presence of systematic undercoverage of the household types are studied through some simple models for coverage. Comparisons with the principal person method are made and the paper concludes with the observation that it is necessary to know more about the nature of survey undercoverage before deciding on which of the constrained minimum distance or principal person methods is to be preferred in applications.

KEY WORDS: Weighting; Auxiliary information; Raking ratio estimation; Principal person method; Survey coverage.

1. INTRODUCTION

Post-stratification is commonly used to adjust survey weights to take into account independent information about the number of units of certain kinds in the population. For example, independent estimates of the population in various age \times race \times sex post-stratification cells may be available from adjusting census counts for known changes in the number of persons since the census. These independent estimates are often referred to as "control counts". Prior to post-stratification, each sample person (or household) has an initial weight, typically corresponding to the inverse of the selection probability. A post-stratification ratio adjustment factor is applied to the weights of all sample persons in each cell, so that the sum of the adjusted person weights equals the independent control count for the cell. This adjustment is especially important when there is systematic undercoverage of households or persons within households.

For most U.S. Census Bureau demographic surveys, post-stratification is used in assigning weights to sample persons, but is not used directly in assigning weights to sample households. This is due to the greater difficulty of obtaining independent estimates for households. Instead, household weights for these surveys are assigned using some version of the "principal person" method. In the basic principal person method, the household weight is set equal to the final post-stratified person weight of the "principal" person in the household. The rule for identifying this person will be described in Section 2. By using the post-stratified person weight, the principal person method does incorporate the independent estimates of persons into the weights assigned to households.

The most obvious problem with the principal person method is that when the resulting household weights are used to calculate weighted estimates of the number of persons in each post-stratification cell, with each person being given his or her household's weight, these

¹Charles H. Alexander, Statistical Methods Division, U.S. Census Bureau, Washington, D.C. 20233 U.S.A.

estimates do not agree with the control counts used in the post-stratification. Consequently, there has been interest in methods of assigning weights to households which are constrained to produce person estimates which agree with the independent control counts.

This paper considers a class of methods for assigning survey weights to households, constrained to be consistent with the "known" control counts in various person cells. The general idea is to find household weights which satisfy the constraints and are as close as possible to the initial vector of weights assigned to the households. The different methods within the class correspond to different ways of measuring the distance between the initial vector of weights and the adjusted vector of weights.

Section 2 describes six "constrained minimum distance" weighting methods of this type plus a version of the principal person method. Three of the six methods have been investigated previously, and the others are added in this paper to round out the picture. Section 3 describes the computation of the weights. Section 4 discusses how the adjusted weight depends on the composition of the household. Section 5 discusses results and examples which may help in understanding what these methods do. Section 6 describes areas for further research.

This work has numerous antecedents. The general class of constrained minimum distance methods is suggested for household weighting by Luery (1986). Extending Luery's work, Zieschang (1986a) proposes using one of these methods, generalized least squares, for weighting the U.S. Consumer Expenditure Surveys. Another member of the class is the "minimum discriminant information method", otherwise known as raking ratio estimation or, simply, raking. Oh and Scheuren (1978a) specifically discuss the raking approach to the household weighting problem, and give additional references to a rich literature on raking and related methods. The idea of viewing raking as a constrained minimum distance problem dates back at least to Deming and Stephan (1940). The fundamental principles of this approach are explored in Ireland and Kullback (1968). Applications to survey weight adjustment are well covered in Brackstone and Rao (1979). The class of methods also includes two criterion functions related to multinomial maximum likelihood. The relationship of this to raking has been extensively studied; see, for example, Bishop, Fienberg, and Holland (1976). Fienberg (1986) points out that the distance criteria considered in this paper may be viewed as special cases of a parametric family of functions considered in Cressie and Read (1984).

2. CONSTRAINED MINIMUM DISTANCE METHODS

2.1 Methods Based on Household Weights

Consider a sample of K households, whose initial weights are given by the vector $\underline{S} = (S_1, \dots, S_K)'$. In this paper, S_k will be the inverse of the probability of selection of the k -th household; in some applications other adjustments such as nonresponse factors may be included in the initial weight.

Suppose that there are J post-stratification cells, and that the number of persons in the population (N_j) is known for each cell. For example, for the U.S. Consumer Expenditure Survey, there are $J = 48$ cells corresponding to combinations of the two sexes, two race (black, nonblack), and twelve age categories. In that survey, persons younger than 1 are not included. The control counts for these cells will be treated as a vector $\underline{N} = (N_1, \dots, N_J)'$.

The composition of the sample households will be described by a matrix $A = (a_{kj})$, where a_{kj} is equal to the number of persons in the k -th sample household who are in the j -th post-stratification cell. Summing over the post-stratification cells for the k -th household gives $a_{k\cdot}$, the total number of persons in the k -th household. For household k , the vector

(a_{k1}, \dots, a_{kJ}) describes the composition of the household. For example, if the vector is $(2, 1, 0, 0, \dots, 0)$, then the household contains exactly two persons in the first cell and one in the second.

Using the initial weights \underline{S} , the weighted sample estimate of the number of persons in cell j would be $\hat{N}_j = \sum_k a_{kj} S_k$ or in general $\hat{\underline{N}} = \underline{A}' \underline{S}$.

Typically $\hat{\underline{N}} \neq \underline{N}$, i.e., the initial weighted estimate of persons in the post-stratification cells may not equal the known population of the cell.

The goal is to define a new vector of weights $\underline{W} = (W_1, \dots, W_K)'$ for the sample households, so that $\underline{N} = \underline{A}' \underline{W}$ or

$$\sum_k a_{kj} W_k = N_j \text{ for } j = 1, \dots, J. \quad (1)$$

The solution to (1) is not necessarily unique. The idea of the constrained minimum distance methods is to choose \underline{W} so as to minimize some measure $D(\underline{W}, \underline{S})$ of the distance between the vectors \underline{W} and \underline{S} , subject to (1). In this way, the initial weights \underline{S} are changed as little as possible in meeting the constraint that the adjusted weights should agree with the known control totals. Note that, for certain possible values N_1, \dots, N_J , it may be impossible for any vector of weights \underline{W} to satisfy the constraints (1). Practically speaking, this possible infeasibility does not seem to be a problem, provided the sample is large enough to include a good representation of different types of households, since the controls \underline{N} are generated from the actual population and therefore can be expected to be "feasible".

There are numerous ways of measuring the difference between two vectors. Three distance criteria $D(\underline{W}, \underline{S})$ will be considered, corresponding to a household-level generalized least squares (GLS-H) objective function, a minimum discriminant information (MDI-H) function, and a maximum likelihood estimation (MLE-H) criterion. The criteria are:

$$\text{GLS - H:} \quad \sum_k (W_k - S_k)^2 / S_k, \quad (2a)$$

$$\text{MDI - H:} \quad (\underline{S} - \underline{W}) + \sum_k W_k \ln(W_k / S_k), \quad (2b)$$

$$\text{MLE - H:} \quad (\underline{W} - \underline{S}) - \sum_k S_k \ln(W_k / S_k). \quad (2c)$$

Throughout the paper, the dot notation is used to denote summation over a subscript.

In each case $D(\underline{W}, \underline{S})$ is nonnegative and is equal to zero if and only if $\underline{W} = \underline{S}$. This can be shown, in the usual way, by examining the first and second partial derivatives of each expression with respect to the W_k .

Algorithms for calculating \underline{W} to minimize these three criteria, while meeting the constraint (1) to the degree of approximation desired, will be discussed in Section 3.

2.2 Methods Derived from Person Weights

An alternative approach to this problem leads to a slight but important modification of the three distance criteria. These modified criteria are given by (5a), (5b), and (5c) below. Although these criteria lead to weights for households, they are generated by an approach which starts out by trying to define weights for persons. Accordingly, first consider the problem as one of defining person weights as close as possible to their original household weights, subject to the constraint that the weighted estimate of persons in each post-stratification cell

equals the known control. Let the persons in the k -th household be numbered $i = 1, \dots, a_k$, and let S_{ki} be the initial weight of the i -th person in the k -th household; note that $S_{ki} = S_k$.

Let b_{kij} be a zero-one indicator variable showing whether the i -th person in the k -th household is in the j -th post-stratification cell. Then the condition for consistency with the controls is

$$\sum_k \sum_i b_{kij} W_{ki} = N_j. \quad (3)$$

The three criteria for the person weighting problem would be

$$\sum_k \sum_i (W_{ki} - S_{ki})^2 / S_{ki}, \quad (4a)$$

$$S_{..} - W_{..} + \sum_k \sum_i W_{ki} \ln (W_{ki} / S_{ki}), \quad (4b)$$

$$W_{..} - S_{..} - \sum_k \sum_i S_{ki} \ln (W_{ki} / S_{ki}). \quad (4c)$$

These criteria could be used for defining person weights. In fact the criterion (4c) would lead to the post-stratification weights which are used in person weighting for the Consumer Expenditure Survey, as described in Alexander (1986). However, our problem is to define weights for households. Household weights may be obtained from these criterion functions by imposing upon the person problem the additional constraint that all persons in the same household must have the same weight. Therefore, let $W_{ki} = W_k$ for $i = 1, \dots, a_k$. Under this constraint, (3) becomes

$$N_j = \sum_k \left(\sum_i b_{kij} \right) W_{ki} = \sum_k a_{kj} W_k,$$

which is the same as the constraint (1) in Section 2.1. The distance criteria (4a), (4b), and (4c) now become:

$$\text{GLS-P:} \quad \sum_k a_k (W_k - S_k)^2 / S_k, \quad (5a)$$

$$\text{MDI-P:} \quad \sum_k a_k S_k - \sum_k a_k W_k + \sum_k a_k W_k \ln (W_k / S_k), \quad (5b)$$

$$\text{MLE-P:} \quad \sum_k a_k W_k - \sum_k a_k S_k - \sum_k a_k S_k \ln (W_k / S_k). \quad (5c)$$

The criteria are now summations at the household level, but the household size a_k has been brought into the criterion for measuring the distance between the initial and adjusted vector of weights. These criteria will be seen to have advantages over the more direct approach which led to (2a), (2b), and (2c).

2.3 The Principal Person Method

In the basic principal person method, the post-stratified person weight of the household's "principal person" is used as the household's weight. To determine the principal person, it is first necessary to determine the household's "reference person". The reference person is identified by the interviewer as the first person mentioned in response to the instruction "start by giving me the name of someone who owns or rents this house." Household relationships are defined in terms of the other members' relationship to this reference person. "Reference person" has replaced the "head of household" concept for this purpose.

The principal person is the wife of the reference person if the reference person is a married male with spouse present. Otherwise, the principal person is the reference person himself or herself. The rationale for this choice is that the principal person should be a person who is not likely to be missed due to within-household undercoverage. In general, women have better coverage than men. Further, the principal owners or renters of the house or apartment seem unlikely to be overlooked.

The basic idea of the principal person method is that there is exactly one principal person in each household. Consequently, the number of households may be estimated by estimating the number of principal persons. This basic method is used for the U.S. National Crime Survey. Other surveys such as the U.S. Consumer Expenditure Surveys or Current Population Survey, make additional adjustments based on assumptions about within-household undercoverage of principal persons, as compared to other persons in the same post-stratification cell (Alexander 1986.)

The principal person method is difficult to model theoretically because the designation of the reference person is somewhat arbitrary. In the hypothetical examples of Section 5, a simplified version of the principal person method will be used, in which the principal person is the household member whose post-stratification cell has the best coverage, i.e., whose post-stratification factor is closest to one. A similar idea is used in Scheuren (1981).

This simplified principal person method will be represented symbolically as follows. For the k -th sample household, let $j(k)$ be the post-stratification cell of the household's principal person. Then the household's principal person weight is

$$W_k = S_k(N_{j(k)} / \hat{N}_{j(k)}).$$

3. COMPUTATION OF THE WEIGHTS

The two least squares methods, GLS-H and GLS-P, have closed-form expressions for \underline{W} , providing that there exists some solution to the constraints (1). For the GLS-H weights, the adjusted weights are given by

$$\underline{W} = \underline{S} + MA(A'MA)^{-1}(\underline{N} - A'\underline{S}) \quad (6)$$

where $\underline{S} = (S_1, \dots, S_K)$, $\underline{N} = (N_1, \dots, N_J)$, A is the matrix (a_{kj}) and M is the $K \times K$ diagonal matrix with the elements of \underline{S} on the main diagonal. The weights \underline{W} for the GLS-P method are also given by (6), except that M is the $K \times K$ diagonal matrix with the values $S_1/a_1, \dots, S_K/a_K$ on the main diagonal.

A disadvantage of (6) for either method GLS-H or GLS-P is that the solution \underline{W} may include negative weights. Conceptually this is unsettling, and for practical users negative weights are unacceptable. It is usually possible to incorporate additional constraints that the

weights must be positive. Ways of doing this are given by Zieschang (1986a) and Huang and Fuller (1978). However, the advantage of a simple closed-form solution is lost with these additional constraints.

The raking method (MDI-P) has been used before for household weighting, e.g., by Oh and Scheuren (1978a). A related method which has been extensively tested is described in Pugh, Tyler, and George (1976), based on the approach of Stephan (1942). Luery (1986) gives an iterative algorithm based on Darroch and Ratcliff (1972), which is proved to converge whenever there is a solution to (1). This method is presented here, since the iterative step has a simple interpretation. The iteration starts with "step 0" weights

$$W_k(0) = S_k(N_{\cdot} / \hat{N}_{\cdot})$$

In other words, the initial weight S_k is adjusted by an overall inflation factor equal to the known population N_{\cdot} divided by the initial weighted total population. At subsequent iterative steps, the adjustment is

$$W_k(i) = W_k(i-1) \prod_j \left(N_j / \sum_s a_{sj} W_s(i-1) \right)^{a_{kj}/a_k}.$$

Note that $W_k(i-1)$ is multiplied by the geometric mean of the post-stratification factors for the persons in the k -th household, where the post-stratification factors are calculated using the weights after iteration $i-1$.

The other three methods, MDI-H, MLE-H, and MLE-P, have not been extensively studied. The following iterative algorithms have worked successfully in small hypothetical examples such as those given in Section 5. In each case, a system of equations, which the weights must satisfy in order to minimize the distance criterion subject to the constraints, can be found by the use of Lagrange multipliers. The equations cannot be solved directly, but if an iterative method produces solutions of the proper form, then the solution minimizes the criterion. If the algorithms converge, the solutions will satisfy the equations. However, the author has no general proof of convergence. A possible alternative approach for the "maximum likelihood" criteria would be to apply the approach of Haber and Brown (1986). Other related work is Fagan and Greenberg (1985).

3.1 Method for MDI-H

The equation for the weights is

$$W_k = S_k \prod_j \gamma_j a_{kj} \quad (7)$$

subject to (1). If values $\gamma_1, \dots, \gamma_J$ can be found so that the weights calculated according to (7) satisfy (1), then those weights minimize (2b) subject to (1). An iterative algorithm for generating such a vector \underline{W} is as follows.

Initialize $W_k(0) = S_k$ and $\gamma_j(0) = 1$. Then at the i -th iteration let

$$\gamma_j(i) = \gamma_j(i-1) \left[1 - (\hat{N}_j(i-1) - N_j) / \sum_s a_{sj}^2 W_s(i-1) \right],$$

where $\hat{N}_j(i-1) = \sum_s a_{sj} W_s(i-1)$. Then let $W_k(i) = S_k \prod_j (\gamma_j(i))^{a_{kj}}$.

3.2 Method for MLE-H

The solution is of the form:

$$W_k = S_k / \left(1 + \sum_j \gamma_j a_{kj} \right).$$

subject to (1).

An iterative solution is

$$W_k(0) = S_k \quad \text{and} \quad \gamma_j(0) = 0,$$

$$\gamma_j(i) = \gamma_j(i-1) + (\hat{N}_j(i-1) - N_j) / \left(\sum_s (a_{sj} W_s(i-1))^2 / S_k \right),$$

$$W_k(i) = S_k / \left(1 + \sum_j \gamma_j(i) a_{kj} \right).$$

3.3 Method for MLE-P

The solution is of the form:

$$W_k = S_k / \left(\sum_j \gamma_{kj} a_{kj} / a_{k.} \right).$$

subject to (1).

An iterative solution is

$$W_k(0) = S_k \quad \text{and} \quad \gamma_j(0) = 1,$$

$$\gamma_j(i) = \gamma_j(i-1) \hat{N}_j(i-1) / N_j,$$

$$W_k(i) = S_k / \left(\sum_j \gamma_j(i) a_{kj} / a_{k.} \right).$$

4. THE ROLE OF A HOUSEHOLD'S "COMPOSITION TYPE"

For the six constrained minimum distance methods, the ratio of a household's initial weight to its adjusted weight depends on the number of people in the household in the different post-stratification cells. To discuss this further, the notion of a household's "composition type" will be introduced. Two sample households, say k and m will be said to "have the same type" if they have exactly the same number of people in each of the post-stratification cells, i.e., if

$$a_{kj} = a_{mj} \text{ for } j = 1, \dots, J. \quad (8)$$

As an example, one household type would be a "household consisting of a white male 35-39 and a white female 30-34." Note that the composition type does not depend on family relationships.

The ratio of the adjusted weight to the initial weight, W_k / S_k , is the same for all households with the same type. In other words, if k and m satisfy (8), then $W_k / S_k = W_m / S_m$. This fact was used in Ireland and Scheuren (1975). A formal proof is given in Alexander and Roebuck (1986).

A useful consequence of this fact is that, in calculating the weights for the constrained minimum distance methods, the calculations may be done using the household type as the unit of analysis rather than the individual household. A simple example may make the implications of these results clearer. Suppose that there are two post-stratification cells, $j = 1$ for females and $j = 2$ for males. The sample consists of K households. For household k the vector (a_{k1}, a_{k2}) describes how many females and males are in the household; household with vector $(2,1)$ has two females and one male.

Practically speaking, there is some upper limit on the size of a household, and there are only finitely many household types. For the example, assume that no household has more than three people. Then there are $T = 9$ household types corresponding to the vectors: $(1,0)$, $(0,1)$, $(2,0)$, $(1,1)$, $(0,2)$, $(2,1)$, $(1,2)$, $(3,0)$, $(0,3)$. These types will be numbered consecutively $t = 1, \dots, 9$. The types will also be labelled mnemonically, F, M, FF, FM, MM, FFM, FMM, FFF, MMM. Hypothetical sample data and control totals are given in Table 1. Note that S_t is the total initial weight given to households of type t .

The constrained minimum distance adjustments effectively may be calculated from the total weights for the household composition types, S_1, \dots, S_9 , without actually looking at the individual household weights. Adjusted weights W_1, \dots, W_9 may be calculated using the algorithms from Section 3 replacing summation over k by summation over t . Then for any type t household, the adjusted weight given by the method is W_t/S_t times the initial weight for the household. (The potentially confusing notation of using S_k for the household weight and S_t for the total weight for a t household type is adopted to emphasize that the formulae of Sections 2 and 3 apply equally well to households or household types. In doing calculations, the meaning will be clear from the context.)

The reduction of the problem from individual households to household types is extremely convenient for presenting small examples. Even when applied to the full 48 post-stratification cells, the household-type approach may still be practical: despite the astronomical number of possible household types, the actual number of types in the sample can never be larger than the sample size and often is substantially smaller. This was found to be the case for related cells of households in Ireland and Scheuren (1975). Simply reducing the size of the computational task by combining the weights for single-person households of the same type may be useful; this has been done at the U.S. Bureau of Labor Statistics in applying the generalized least squares method to the Consumer Expenditure Surveys.

The simplified version of the principal person method also depends only on the household type. If two households have the same composition, then their principal persons will be in the same post-stratification cell, the one with the post-stratification factor closest to one. Consequently, the same ratio adjustment factor would be used for both households. In the actual principal person method, the principal person depends in part on who happens to be designated as reference person, so the adjustment factor is not completely determined by the household's composition type.

Note that the MLE-H method corresponds to calculating multinomial maximum likelihood estimates (subject to the constraint (1)) of p_t , $t = 1, \dots, T$, where p_t is the population proportion of households with type t . The MLE-P method has a related interpretation. Neither of these models, which also pertain to the corresponding GLS and MDI methods, allow for systematic undercoverage.

5. DISCUSSION OF THE METHODS

This Section begins with some speculations about properties of the constrained minimum distance methods, based on the results of Section 4, and follows with some simple hypothetical examples, which generally appear to support the speculations.

The first conjecture is that MLE-H, GLS-H, and MDI-H will tend to give similar results, and also that MLE-P, GLS-P, and MDI-P will tend to be similar to one another, at least for large samples. This is based on the observation that these are all best asymptotic normal estimators under the relevant multinomial sampling model, where the cells are the household types. For small or moderate sample sizes, greater differences between the methods might be anticipated, especially if there are a large number of household composition types, so that the sample in individual "cells" of the multinomial may be small.

The examples given below tend to support this conjecture; the "household" methods all give very similar results, as do the "person" methods. This is true even in some cases when the hypothetical data do not fit the model very well. However, these examples involve only a small number of household types and post-stratification cells, and so are illustrative rather than conclusive.

The second conjecture is based on considering the nature of the sampling models under which the constrained minimum distance methods may be viewed as maximum likelihood estimates, or asymptotic approximations thereto. In these models, perfect coverage is assumed. The models assume a distribution corresponding to probabilities which are the actual proportions in the population, and these probabilities are consistent with the "true" control totals used in the constraints (1). According to these models, for sufficiently large samples, the initial sample estimates would approach agreement with the control totals. This would not be true when there is substantial undercoverage in the sampling frame. Such undercoverage is an important reason for using post-stratification. Coverage considerations may be especially important for telephone surveys where there is no supplemental frame to include households without telephones. If there is no special adjustment for noninterview "nonresponse", such as refusal or inability to provide the requested information, then nonresponse may be a further departure.

Based on these remarks, the second conjecture is that without adjustment the constrained minimum distance methods may not perform well in adjusting for systematic undercoverage, even for large samples. The methods are optimal under models which assume perfect coverage; one would expect that they might be less than optimal when this assumption is violated.

The examples given below partly support this conjecture. The constrained distance methods do not do as well as the simplified principal person method under certain assumptions about undercoverage. Under other assumptions, some of the methods may do quite well. The author concludes that it is necessary to know more about the nature of survey undercoverage before judging that any of these methods is superior to the principal person method. Oh and Scheuren (1978b) raise some related issues about mean square error of the raking estimator when there is undercoverage.

Two examples will be presented, representing two extreme forms of undercoverage. The first ("household undercoverage example") will assume that there is a uniform 10% undercoverage of all households, but that there is no within-household undercoverage. The second example ("within-household undercoverage example") assumes a 10% undercoverage of males due to within-household undercoverage in households where there are both males and females, and undercoverage of all-male households. For single-person households, any "within-household undercoverage" means that the whole household is missed.

In example 1, there is a 10% under-representation of all types of households in the sample. For a sufficiently large sample, this would obviously be due to systematic undercoverage, rather than sampling error. Applying the constrained minimum distance methods and the principal person method to this example gives the total adjusted weights for each household type shown in the last four columns of Table 1.

Note that the GLS-P, MDI-P, and MLE-P methods all bring the adjusted weight up to the actual population value. Thus, these methods give "unbiased" weights. Since all persons have a second-stage factor of $1/.9$, the principal person method also achieves this result.

Table 1
Household Undercoverage Example:
Description of Population and Sample

Type & description	Actual Population	Total Initial Weights	Total Weight (W_i) for Methods:			
			GLS-H	MDI-H	MLE-H	GLS- MDI-H MLE-H Prin. Pers.
1: F	25,000	22,500	23,785	23,745	23,704	25,000
2: M	15,000	13,500	14,120	14,097	14,075	15,000
3: FF	7,000	6,300	7,020	7,016	7,013	7,000
4: FM	40,000	36,000	39,708	39,672	39,632	40,000
5: MM	5,000	4,500	4,913	4,906	4,900	5,000
6: FFM	12,000	10,800	12,529	12,506	12,594	12,000
7: FMM	12,000	10,800	12,408	12,428	12,449	12,000
8: FFF	0	0	0	0	0	0
9: MMM	0	0	0	0	0	0
Total	116,000	104,400	114,483	114,370	114,367	116,000
Control Totals:		Number of Females	=	115,000		
		Number of Males	=	101,000		
Initial Weighted		Females	=	103,500		
Person Counts:		Males	=	90,900		

The other methods, GLS-H, MDI-H, and MLE-H, all give substantially too little weight to one-person households and too much to the three-person households. Intuitively, this makes sense; since these methods do not allow for systematic undercoverage and must explain the shortage of sample persons as sampling error, the obvious explanation is that the sample has a below-average number of large households, due to chance. The better performance of MLE-P makes some sense, since it starts out with a multinomial sampling model which allows sampling of persons without regard to households.

Practically speaking, this example reflects very poorly on the GLS-H, MDI-H, and MLE-H methods. Even uniform undercoverage would cause these methods to distort the distribution of household sizes. Worse, the distortion goes opposite from what is commonly assumed about differential household coverage, namely that small households are more likely to be missed than large ones, so that small households need relatively higher weights, not relatively lower weights.

The second example will emphasize within-household undercoverage of males. The situation is more complicated than in the previous example, because a household may have an apparent composition type different than its actual type. For example, a household which actually consists of a male and a female may appear to be a single-person household. The actual and apparent type will be indicated by modifying our previous notation. For example, a FM household in which the male is missed will be denoted F|M|. A |M| household or [MM] household is missed entirely. Table 2 describes the hypothetical data. The actual population is the same as in the previous example.

Table 2
Within-household Undercoverage Example:
Description of Population and Sample

Actual Household Type	Apparent Type	Actual Number	Total Initial Weights
1: F	F	25,000	25,000
2: M	M	13,500	13,500
	[M]	1,500	0
3: FF	FF	7,000	7,000
4: FM	FM	36,000	36,000
	F[M]	4,000	4,000
5: MM	MM	4,500	4,500
	[MM]	500	0
6: FMM	FFM	10,800	10,800
	FF[M]	1,200	1,200
7: FMM	FMM	10,800	10,800
	FM[M]	1,200	1,200
8: FFF	FFF	0	0
9: MMM	MMM	0	0
		116,000	114,000
Control Counts:	Number of Females	115,000	
	Number of Males	101,000	
Initial Weighted Person Counts:	Females	115,000	
	Males	90,900	

Note that there is a 10% undercoverage of males, due to missing males within households, or missing all-male households. Each male has a 10% chance of being missed.

Neither column of numbers in table 2 is observed, since there are no household controls. Also the actual household type is not known for the sample units. Thus, the [FM] households appear to be the same as the F households. The data which would be observed are given in Table 3, along with the total initial weight for households which appear to have a given type. The adjusted weights are given for three methods, MLE-H, MLE-P, and principal person. The results for GLS-H and MDI-H are fairly close to MLE-H, and GLS-P and MDI-P are similar to MLE-P, so these other methods are omitted.

The last three columns of Table 3 show the total adjusted weight assigned to each actual household type by the MLE-H, MLE-P, and principal person methods. The principal person weights for each actual household type agree with the population counts for the actual types, shown in the third column of Table 1. In this sense, the principal person weights are unbiased.

This example corresponds to assumptions upon which the simplified principal person is based. The principal person adjusted weights for each actual type of household coincide with the population counts. The one difference is that totally missing [M] or [MM] households are given no weight; however, the weight of the non-missing M or MM households is increased accordingly. The total weighted number of households for the principal person method is equal to the number in the population.

Table 3
 Within-household Undercoverage Example: Observed Types and Weights,
 with Adjusted Weights from Three Methods

Household Type	Total Initial Weight	Weight Assigned to Apparent Type			Weight Assigned to Actual Type		
		MLE-H	MLE-P	Principal Person	MLE-H	MLE-P	Principal Person
F	29,000	27,450	26,973	29,000	23,664	23,253	25,000
M	13,500	14,997	16,338	15,000	14,997	16,338	15,000
FF	8,200	7,368	7,626	8,200	6,290	6,510	7,000
FM	37,200	38,887	39,128	37,200	41,419	41,586	40,000
MM	4,500	5,623	5,446	5,000	5,623	5,446	5,000
FFM	10,800	10,661	10,885	10,800	11,739	12,001	12,000
FMM	10,800	12,605	11,878	10,800	13,859	13,140	12,000
FFF	0	0	0	0	0	0	0
MMM	0	0	0	0	0	0	0
Total	114,000	117,591	118,274	116,000	117,591	118,274	116,000

In this example, the constrained minimum distance methods overestimate the total number of households, but give too little weight to the households without males. In general, too much weight is given to households with males.

It should not be concluded that the principal person method always outperforms the constrained minimum distance methods when there is within-household undercoverage. Under other assumptions about coverage, the principal person method may not do so well. In fact, different versions of the principal person method are used for different surveys, based on various assumptions about coverage. Note also that combinations of the principal person method and raking methods are possible; see Scheuren (1981).

Even in this example, the biased weights assigned by the constrained minimum distance methods could be beneficial for estimating some characteristics. If the households in which males are missed tend to under-report the variable of interest, then giving these households too high a weight may tend to counteract response bias associated with the within-household undercoverage.

The most extreme example of this effect is estimation of the total number of males, in which case the MLE-H and MLE-P weights give estimates which agree with the control totals while the principal person weights do not. However, for household characteristics where the males would rarely be reporting errors because of the missed male, such as form of tenure (renter/owner), the biased weights would not be desirable. The performance of the weighting methods in situations like these clearly depends on the nature of the survey undercoverage and its relationship to the variable being estimated. This is discussed further, with additional examples, in Alexander and Roebuck (1986).

Pending further research on survey coverage and its effect on weighting, what recommendations can be made? Among the constrained minimum distance methods considered in this paper, GLS-H, MDI-H, and MLE-H seem unattractive because of their failure to adjust correctly for uniform undercoverage of households. This is in spite of the fact that, if there were no undercoverage, MLE-H seems to be based on a more sensible model than MLE-P, since households rather than persons are the ultimate sampling unit.

The possibility of negative weights raises questions about the appropriateness of GLS-P, even though in some practical applications (such as Zieschang 1986b) there are very few negative weights, so that they could be replaced by positive weights with little effect on the estimates. That leaves MDI-P and MLE-P. Our results give little basis for choosing between these methods. Computational considerations tend to favor the "raking" method MDI-P. Based on limited experience with the algorithms of Section 3, the MLE methods converge more slowly than the MDI methods. Further, there has been considerable research into ways to improve the efficiency of raking for large-scale applications, such as Ireland and Scheuren (1975). Taking all this into account, the raking method, MDI-P, seems to be the most promising of the constrained minimum distance methods.

The constrained minimum distance methods give household weights which are consistent with control totals for person, unlike the principal person method. However, the superiority of the constrained minimum difference methods over the principal person method as an adjustment for undercoverage is far from obvious. Undercoverage is an essential part of the survey weighting problem. The principal person method is an ad hoc solution to the undercoverage problem, based on some very simplistic assumptions about coverage. However, as seen in Section 4, the constrained minimum difference methods may be viewed as "optimal" (i.e., maximum likelihood or the asymptotic equivalent) estimators under models which assume perfect coverage. The choice is thus between an optimal solution to the wrong problem and an ad hoc solution to what may or may not be the right problem. Clearly more research is needed.

6. SOME AREAS FOR FURTHER RESEARCH

6.1 Household Control Totals

If independent estimates of the number of households of different kinds were available, then ordinary post-stratification could be used for household estimates. Household controls by size of household are being investigated, based on updating 1980 census results (Das Gupta *et al.* 1986). The availability of household controls would fundamentally change our ability to deal with the household weighting problem.

Even with household controls, it might be beneficial to also incorporate person controls. The household controls are not likely to include detailed information on the age, race, and sex of the household members. The use of raking to simultaneously control the estimates to independent controls for persons and households is developed by Scheuren (1981), using an estimate of the total number of households. Zieschang (1986a) describes how similar adjustments may be made using generalized least squares.

Household controls clearly have great potential for adjusting for differential coverage of various types of households. There still may be problems in dealing with within-household undercoverage, since this may lead to errors in determining the true household size, which would cause sample households to be placed in the wrong post-stratification cell.

6.2 Research Concerning Coverage

Coverage of persons is measured fairly well by comparing the initial survey estimates \hat{N}_j to the control totals N_j . It is difficult to determine how much of this undercoverage is due to missing entire households and how much is due to missed persons within households. Additional information could be obtained by comparing initial weighted household estimates

to household controls, once these controls become available. In the meantime, 1980 survey estimates by type of household could be compared to the corresponding 1980 census counts.

Even with this additional information, it is not possible to completely distinguish household undercoverage from within-household undercoverage, without making additional assumptions. Alexander and Roebuck (1986) present some preliminary suggestions about how a range of coverage models might be fit to census and survey data. An alternative approach would be to include coverage parameters in a multinomial sampling model such as those described for the MLE-H or MLE-P weighting methods. Other approaches to modelling coverage are presented in Wolter (1986).

6.3 Estimation of Variances

Methods for estimating variances of the weighted estimators have not been investigated for most of the constrained minimum distance methods. For raking estimators, some methods are available; see Arora and Brackstone (1977), Bankier (1978) and Fan *et al.* (1981).

For any of the methods, replication methods for estimating the variance could be applied. These methods have been shown to give reasonable results under fairly general conditions; see for example Krewski and Rao (1985). It remains to be determined whether these conditions can be applied to the constrained minimum distance methods.

6.4 Computational Issues

Zieschang (1986b) has applied the generalized least squares methods to the U.S. Consumer Expenditure Surveys. Scheuren (1981) describes a large-scale application of the raking method to household weighting. The maximum likelihood constrained minimum distance algorithm (MLE-H and MLE-P) have not been tried on large-scale problems of this kind. If they were to be used in actual survey weighting, research may be needed to improve their computational efficiency.

ACKNOWLEDGMENTS

The author would like to thank Michael J. Roebuck for his assistance with portions of this research, and also the associate editor and referees for their helpful comments. The author is also indebted to Brenda Kelly for her diligence in typing this manuscript.

REFERENCES

- ALEXANDER, C.H. (1986). The present Consumer Expenditure Surveys weighting method. In *Population Controls in Weighting Sample Units*, Section 1. Washington, D.C.: U.S. Bureau of Labor Statistics, 1-32.
- ALEXANDER, C.H., and ROEBUCK, M.J. (1986). Comparison of alternative methods for household estimation. *Proceedings of the Section on Survey Research, American Statistical Association*, 54-57.
- ARORA, H.R., and BRACKSTONE, G.J. (1977). An Investigation of the Properties of Raking Ratio Estimates: II. With cluster sampling. *Survey Methodology*, 4, 232-252.
- BANKIER, M.D. (1978). An estimate of the efficiency of raking ratio estimators under simple random sampling. *Survey Methodology*, 4, 115-124.
- BRACKSTONE, G.J., and RAO, J.N.K. (1979). An investigation of raking ratio estimators. *Sankhyā Series C*, 41, 97-114.

- SHOP, Y.M.M., FIENBERG, S.W., and HOLLAND, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, Massachusetts: MIT Press.
- RESSIE, N., and READ, T.R.C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society, Series B*, 46, 440-464.
- ARROCH, J.N., and RATCLIFF, D. (1972). Generalized iterative scaling for log-linear models. *Annals of Mathematical Statistics*, 63, 1470-1480.
- AS GUPTA, P., GIBSON, C., HERRIOT, R.A., LAMAS, E., and ZITTER M. (1986). New approaches to estimating households and their characteristics for states and counties. Paper presented at the 1986 annual meeting of the Population Association of America.
- EMING, W.E., and STEPHAN, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected margins are known. *Annals of Mathematical Statistics*, 11, 427-444.
- AGAN, J.T., and GREENBERG, B. (1985). Algorithms for making tables additive: raking, maximum likelihood, and minimum chi-square. U.S. Bureau of the Census, Statistical Research Division Report Series No. Census/SRD/RR-85/12.
- AN, M.C., WOLTMAN, H.F., MISKURA, S.M., and THOMPSON, J.H. (1981). 1980 census variance estimation procedure. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 176-181.
- ENBERG, S.E. (1986). Comments on some estimation problems in the Consumer Expenditure Surveys. In *Population Controls in Weighting Sample Units*. Section 5. Washington, D.C.: U.S. Bureau of Labor Statistics, 1-12.
- ABER, M., and BROWN, M.B. (1986). Maximum likelihood methods for log-linear models when expected frequencies are subject to linear constraints. *Journal of the American Statistical Association*, 81, 477-482.
- UANG, E.T., and FULLER, W. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of Social Statistics Section, American Statistical Association*, 300-305.
- ELAND, C.T., and SCHEUREN, F.J. (1975). The rake's progress, *Computer Programs for Contingency Table Analysis*. Washington, D.C.: The George Washington University, 155-216.
- REWSKI, D., and RAO, J.N.K. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics* 9, 1010-1019.
- JERY, D.M. (1986). Weighting sample survey data under linear constraints on the weights. *Proceedings of the Social Statistics Section, American Statistical Association*, 325-330.
- H, H.L., and SCHEUREN, F.J. (1978a). Multivariate raking ratio estimation in the 1973 Exact Match Study. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 716-722.
- H, H.L., and SCHEUREN, F.J. (1978b). Some unresolved application issues in raking ratio estimation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 723-725.
- UGH, R.E., TYLER, B.S., and GEORGE, S. (1976). Computer-based procedure for N-dimensional adjustment of data - NJUST. U.S. Social Security Administration, Staff Paper No. 24.
- CHEUREN, F.J. (1981). Methods of estimation for the 1973 Exact Match Study. *Studies from Interagency Data Linkages, Report No. 10.*, U.S. Department of Health and Human Services, U.S. Social Security Administration, 9-122.
- TEPHAN, F.F. (1942). An iterative method of adjusting sample frequency tables when expected marginal totals are known. *Annals of Mathematical Statistics*, 13, 166-178.
- OLTER, K.M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, 81, 338-346.

- ZIESCHANG, K.D. (1986a). Generalized least squares: an alternative to principal person weighting. In *Population Controls in Weighting Sample Units*, Section 2. Washington, D.C.: U.S. Bureau of Labor Statistics, 1-41.
- ZIESCHANG, K.D. (1986b). A generalized least squares weighting system for the Consumer Expenditure Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 64-71.

An Integrated Method for Weighting Persons and Families

G. LEMAÎTRE and J. DUFOUR¹

ABSTRACT

Household surveys generally use separate procedures for estimating characteristics of persons and those of families. An integrated procedure is proposed and a least-squares estimator introduced to achieve this end. The estimator is shown to be unbiased under certain general conditions. Using data from the Canadian Labour Force Survey, variances for the estimator are calculated and shown to compare favourably to those from current procedures.

KEY WORDS: Family estimation; Family weighting; Least-squares weighting.

1. INTRODUCTION

It is customary for many household surveys to incorporate in their estimation procedures a post-stratification step in which the design-based estimates of the population, generally by age and sex group, are benchmarked to independent totals obtained from demographic sources. In practice, for ease of tabulation, a weight is normally associated with each responding person, equal to the product of the inverse sampling rate, an adjustment for non-response, and an age/sex ratio adjustment factor. Estimates for a particular characteristic are then obtained by summing up the weights of all responding persons in the sample bearing that characteristic. Because of the age/sex adjustment factors, the weight so assigned will usually differ from person to person within the same household. When estimating characteristics of persons, this may not pose any particular problem; in producing estimates of households or families, however, it is not entirely clear which weight is the appropriate one to use, if any. To estimate family characteristics, one might well elect to carry out a ratio estimation step using auxiliary information on families as well as persons. However, reliable and timely auxiliary counts of families that could be used in ratio estimation are in general not available. As a result of events such as births, deaths, marriages, divorces and persons leaving or entering a household, characteristics such as family size change from one census to the next, in ways that are less predictable than a characteristic such as age. The administrative records that are the main source of information on post-censal population change (i.e. birth, death and migration records), do not provide information on household-related change. Birth records, for example, do not provide information on the size of a family into which a child is born. Tax records can compensate in part for this deficiency (see Auger 1987); however, such records do not cover the entire population nor are they available in a timely enough fashion to be used in producing current estimates. In the absence of auxiliary counts of families, household surveys generally have adapted the weights obtained from "person-weighting" for use in estimating characteristics of families. For various reasons this is a somewhat less than ideal solution. The present paper proposes a method of estimation that results in a single uniquely defined weight per household which would be appropriate for both individual and family estimation.

¹G. Lemaître and J. Dufour, Social Survey Methods Division, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, K1A 0T6.

Techniques to achieve a single household weight have been proposed in the past, with an emphasis on using auxiliary information on persons to improve estimates of families. Oh and Scheuren (1978) proposed a method of "multivariate raking" which consists of successively ratio adjusting population estimates by post-stratum by means of the ratio adjustments calculated for each post-stratum in turn, and then iterating to convergence. The adjustments at each stage are applied to households containing persons in the particular post-stratum being adjusted for. Zieschang (1986) adopted a Generalized Least Squares (GLS) approach in which the sum of weighted squared adjustments to the design weights were minimized, subject to a set of linear constraints. Alexander (1987) examines several constrained minimum distance weighting methods, including the GLS method, and evaluates them in the context of survey undercoverage. Although the above methods were originally proposed as ways of improving estimates of families, the survey weights derived from the various estimators can clearly be used to estimate characteristics of persons as well. This paper argues in favour of adopting such an integrated approach to individual and family estimation. Section 2 discusses the limitations of the current approaches to estimating characteristics of persons and families. Section 3 introduces a model-based estimator adapted from a generalized weighting procedure due to Bethlehem and Keller (1987). Section 4 presents some empirical results taken from the Canadian Labour Force Survey. Section 5 discusses plans for further study.

2. CURRENT ESTIMATION PROCEDURES

The principal mandate of most household surveys traditionally has been to produce estimates for characteristics of persons, particularly of labour force characteristics. Such surveys adopt the household as the ultimate sampled unit essentially for reasons of cost and convenience. Although the household unit is normally respected in preliminary weighting steps (non-response adjustments, rural/urban adjustments, etc.), it is generally ignored in the final weighting step, i.e. no allowance is made for the fact that the members of a household are sampled as a unit. In particular, any coverage biases associated with the sampled unit are not directly taken into account or compensated for in estimation. Undercoverage is thus assumed to be ignorable in the sense of Rubin (1976); every person in an age/sex post-stratum is treated the same in estimation whether he/she is living alone or comes from a multi-person household. One study of non-response in the Labour Force Survey (Paul and Lawes 1987) however, has demonstrated that smaller households, particularly households without children, tend to be underrepresented in the sample. Although no comparable studies exist for missing households in the Labour Force Survey, studies of private household undercoverage in the census have shown that non-enumerated households are indeed smaller on average than enumerated households (Gosselin and Thérault 1980). A missing-at-random type procedure can lead to biases in labour force estimates for persons, particularly if the labour force distribution of persons in smaller households is different from that of persons in larger ones, things being equal. Intuitively, an estimation procedure which takes into account (even only indirectly) the fact that smaller households are more subject to non-response and undercoverage than larger ones could correct in part for this deficiency in the sample.

In the absence of auxiliary information on households or families that could be incorporated into an appropriate weighting procedure to produce a well-defined family weight, many current methods adopt as the family weight the weight of a "principal person" in the family. In the Canadian Labour Force Survey, this person is the female spouse if present, otherwise the head. Since such methods do not take household composition into account

Family estimates generated using this weight tend to overestimate larger families and to underestimate unattached persons. In addition many characteristics (e.g., population, income) can be estimated using either the individual weight or the family weight, and the estimates will in general disagree, sometimes substantially. Of course even under ideal sampling and interviewing conditions, with no differential non-response or undercoverage, family and individual-based estimates of the same characteristic will disagree somewhat. With a large enough sample, however, the discrepancies should be small. Under actual, i.e., less than ideal conditions, differences may be too large to explain away by a facile appeal to sampling variability. An estimation procedure that yields a single household weight which, when used as an individual weight, respects the auxiliary population totals will eliminate the awkwardness of having two estimation systems. It is these deficiencies that the estimator described in the following section was designed to deal with.

3. A PROPOSED ESTIMATOR

We begin by introducing a generalized weighting procedure based on linear models due to Bethlehem and Keller (1987) and applying it first to person-based estimation as was done in their paper. A modification of the procedure is introduced which leads to household weights appropriate for estimating characteristics of persons. We will borrow freely from their original presentation in what follows.

Assume a survey target population consisting of N units, an N -vector Y of values of a target variable, and an N by p matrix X of auxiliary variables defined for each unit of the target population. The population totals for each auxiliary variable are assumed to be known and will be denoted collectively by the p -vector x . In our application x will consist of age-sex totals. If the auxiliary variables are correlated with the target variable, then for an appropriate p -vector B , the values of $E = Y - XB$ will vary less than the values of the target variable Y . Ordinary least squares on all units of the target population yields

$$B = (X'X)^{-1}X'Y, \quad (3.1)$$

provided X is of full rank. A sample-based estimate for B is given by

$$\hat{B} = (X'\Pi^{-1}TX)^{-1}X'\Pi^{-1}TY, \quad (3.2)$$

where T is a diagonal matrix whose i -th element is 1 if the i -th unit of the population is in the sample, 0 otherwise, and $E(T) = \pi$.

It can be shown that for large samples \hat{B} will be approximately unbiased. The parameter of interest, however, is not B but the population total y . If we define $\hat{y} = \hat{B}'x$, \hat{y} will be an approximately unbiased estimator of y provided that $B'x = y$, or equivalently, provided the sum of the residuals for the population model $Y = XB + E$ is equal to zero. This will hold if the N -vector whose elements consist of ones is in the space spanned by the columns of X , and in particular, if the auxiliary variables X include an exhaustive and mutually exclusive set of indicator variables (for age/sex groups, for example).

If we write $\hat{y} = \hat{B}'x = Y'\Pi^{-1}TX(X'\Pi^{-1}TX)^{-1}x$, we see that the estimator implicitly defines an N -vector of weights given by

$$W = \Pi^{-1}TX(X'\Pi^{-1}TX)^{-1}x,$$

that do not depend on the particular target variable being estimated. If these weights are used to produce sample estimates for the auxiliary variable characteristics, we have that $X'W = x$, so that the weights do indeed yield the appropriate population totals. Furthermore if X consists exclusively of an exhaustive and mutually exclusive set of indicator variables then the regression estimator \hat{y} will be equivalent to the ordinary post-stratification estimator. For further details, see Bethlehem and Keller (1987).

The weight of an arbitrary sample person i under this procedure can be expressed generally as

$$W_i = \sum_j \frac{x_{ij}b_j}{\pi_i}, \quad (3.3)$$

where $(b_1, \dots, b_p) = (X'\Pi^{-1}TX)^{-1}x$ and π_i is the inclusion probability for person i . This suggests that the estimation method described above can be adapted to yield the desired weights by defining the auxiliary variables in the same way for all household members. An obvious way to do this is to define auxiliary variables at the household level, for example by replacing the corresponding variables defined at the person level by the household mean. More formally let Z be an N by p matrix defined for person i ($i = 1, \dots, N$) belonging to household h ($h = 1, \dots, H$) by

$$Z_{ij} = \frac{U_{hj}}{n_h},$$

where U_{hj} is the total for characteristic j in household h , i.e. $U_{hj} = \sum_k X_{kj}$, with the summation being over all members k of household h , n_h = size of household h , and $\sum_h n_h = N$. Let Y again be an N -vector of values for an arbitrary target variable defined on persons. As in person-level estimation, we work with the population model $Y = ZC + E$ and apply least squares to the sample data to obtain an estimate

$$\hat{C} = (Z'\Pi^{-1}TZ)^{-1}Z'\Pi^{-1}TY. \quad (3.4)$$

We define $\hat{y} = \hat{C}'x$ where x is again the vector of population totals for the auxiliary variables. \hat{y} will be an approximately unbiased estimator of y provided the N -vector of ones is in the space spanned by the columns of Z . In a manner analogous to (3.3), the weight for an arbitrary sampled person in household h will be given by

$$W_h = \sum_j \frac{U_{hj}c_j}{\pi_h n_h}. \quad (3.5)$$

Since each household member contributes the same row vector to Z and since each has the same first order inclusion probability, each person within a household will have the same weight. Furthermore the use of the household weight as a person weight yields the correct auxiliary population totals. Although it is possible to obtain negative weights under this procedure (if some of the c_j 's are less than zero), for well-behaved samples (i.e., not subject to serious non-response or undercoverage) households whose weights are changed substantially by this procedure tend to be households of unusual composition that are uncommon in the sample and in the population at large. Recently in weighting twenty-four months of Labour Force

Survey data under this procedure, only one household had a (small) negative weight attributed to it. Negative weights are problematic because it is difficult to attach the usual meaning one assigns to weights, that is, the number of persons/households in the population at large represented by a particular sampled person/household. However, under the formulation described above, the final weights are defined only implicitly and indeed could be viewed as merely a convenient means of generating estimates. In practice even with some negative weights, it is unlikely that a meaningful estimate of level for a characteristic of interest would turn out negative. The problem of explaining a negative weight to a mystified user is of course a different question.

The variance of the estimator $\hat{y} = \hat{C}'x$ described in this paper can be obtained using methods described in Fuller (1975). In addition the estimator can be shown to be equivalent to the GLS estimators proposed by Zieschang (1986) and Alexander (1987) when the space spanned by the auxiliary variables Z contains a vector of ones. Further properties of this type of estimator can be found in Wright (1983).

4. EMPIRICAL RESULTS

The Canadian Labour Force Survey is a monthly rotating panel survey of approximately 48,000 households across Canada (see Platek and Singh 1976 and Singh, Drew, and Choudhry 1984). Households once selected remain in the sample for six consecutive months before being replaced. The primary geographic strata are the ten provinces. Sample sizes vary from a low of 1500 households in Prince Edward Island, the smallest province, to about 9000 households in Ontario, the most populous one. The survey collects data concerning the labour market situation of respondents during a reference week each month and publishes a wide variety of estimates related to the nation's labour supply.

A preliminary evaluation of the estimator described above was carried out using data from one of the monthly surveys. May 1981 was chosen to permit comparisons to results from the 1981 census held at about that time. Although we have been using the terms "household" and "family" interchangeably up to now, user interest is often focused on estimates of "economic families", which consist of all persons in a household related by blood, marriage, or adoption. For weighting purposes it is conceptually more appealing to deal with the actual sampled unit, i.e. the household. However, the empirical results presented here will be based on estimates for economic families. The evaluation carried out focused on both characteristics of persons (labour force status) and of families (number of economic families and number of unattached persons). The least-squares weighting was carried out for two sets of five-year age/sex groups, with persons seventy and over being grouped according to sex. The first set of (twenty-four) age/sex groups excluded children 0 to 14 years of age from the weighting, to permit a comparison to a standard person-based post-stratification estimator using the same auxiliary information. The second set included children grouped into six age/sex groups and was used only for least-squares weighting, since under standard post-stratification the weighting of children would have no effect on the weighting of persons 15 and over.

Although all estimators considered are approximately unbiased for estimates of characteristics of persons, each makes different assumptions about the nature of under-coverage and non-response. (The Labour Force Survey's non-response adjustment procedure assumes that non-responding households are missing at random within geographic area). The post-stratification estimator implicitly assumes that any differential non-response and under-coverage depends only on age and sex and is therefore adequately compensated for by

person-based estimation using auxiliary information on these characteristics. Under least squares weighting, the weight of a person will depend on the age/sex composition of the household (without children in one case, with children in the other). Thus, all things being equal, one would expect the design weight of a person belonging to an age/sex group subject to substantial undercoverage to be adjusted less if that person is living with persons belonging to age/sex groups well covered by the sample than if he/she is living alone.

Since the auxiliary population totals by age and sex are available by province, estimation was carried out separately for each province. However, the smaller provinces have been collapsed into two groups in the following tables.

In general the three estimators do not yield substantially different estimates, particularly A and B. The inclusion of children in the weighting does appear to lead to slightly higher estimates of employment and of unattached persons and slightly lower estimates of economic families nationally and in the larger provinces (compare results from Scheuren *et al.* 1981). This is in line with expectations, although there is still some ground to cover vis-a-vis census results, which show (rounded to thousands) 6,369,000 economic families and 2,583,000 unattached persons at the national level. The moral of the tale is that, although the least-square estimator does take us part of the way home (when the presence of children is taken into account), it will require accurate and timely auxiliary information to eliminate the residual bias

Table 1

Number of Persons Employed and Unemployed, Number of Economic Families and Unattached Persons, Labour Force Survey, May 1981 (In Thousands)

Estimator ^a		Employed	Unemployed	Economic Families	Unattached Persons
Canada	A	11,094	850	6,424	2,432
	B	11,090	850	6,446	2,442
	C	11,120	851	6,410	2,495
Atlantic Region	A	819	102	563	156
	B	819	102	570	154
	C	821	102	569	156
Quebec	A	2,725	304	1,723	587
	B	2,724	304	1,725	596
	C	2,735	305	1,714	614
Ontario	A	4,198	274	2,325	863
	B	4,200	273	2,325	861
	C	4,211	273	2,310	881
Prairie Region	A	2,074	83	1,078	506
	B	2,072	84	1,089	510
	C	2,074	83	1,085	517
British Columbia	A	1,277	88	735	319
	B	1,276	88	738	321
	C	1,280	88	734	327

^a A = post-stratification/principal person, B = least squares with children excluded from weighting and C = least squares with children included in weighting.

The expected performance of the least-squares estimator with regard to efficiency is not altogether obvious. Certainly, if one were to base a prediction on the results observed above, then the similarity of the estimates to those produced by the post-stratification estimator would lead one to expect it to perform as well as the latter. On the other hand, one might expect efficiency gains for estimates of economic families, because of the fact that the least-squares estimator makes use of the auxiliary population totals in determining the household weight. However, a single weight per household is not achieved without some redistribution of weights at the micro level.

Table 2

Distribution of Percent Deviations of Final Weights Relative to the Design Weights, Labour Force Survey, May 1981

Percent Deviation	Percentage of Total Sample		
	Post-Stratification	Least-Squares	Least-Squares (With Children)
> -30%	0.0	0.1	0.2
-30 to -20%	0.0	0.5	0.9
-20 to -10%	0.6	3.0	5.3
-10 to 0%	23.9	20.4	27.1
0 to 10%	53.9	44.6	37.3
10 to 20%	20.6	26.3	21.6
20 to 30%	0.6	4.4	6.2
30 to 40%	0.1	0.4	0.9
40 to 50%	0.0	0.0	0.2
< 50%	0.0	0.0	0.2

Note: Sample size is $N = 159014$.

Table 3

Estimated Efficiencies of Least-Squares Estimators Relative to Post-Stratification Estimator, Labour Force Survey, May 1981

Estimator ^a		Employed	Unemployed	Economic Families	Unattached Persons
Canada	B	1.044	0.999	1.565	1.038
	C	1.066	0.999	1.616	1.036
Atlantic Region	B	1.110	0.977	1.266	0.998
	C	1.193	0.992	1.567	1.070
Quebec	B	1.059	1.005	1.553	1.020
	C	1.063	0.992	1.582	0.992
Ontario	B	1.028	1.011	1.825	1.064
	C	1.059	1.010	1.828	1.037
Prairie Region	B	1.001	1.009	1.205	1.009
	C	1.072	1.066	1.420	1.134
British Columbia	B	1.038	0.964	1.248	1.048
	C	1.053	0.978	1.203	1.045

^a B = least squares with children excluded from weighting and C = least squares with children included in weighting.

As Table 2 illustrates, the least-squares weights have a somewhat greater dispersion than those based on standard post-stratification methods. Including children in the weighting results in an even greater dispersion. The movement in the weights essentially reflects the extent to which the age/sex household size composition of the sample fails to mirror that existing in the general population. Since the objective of a single weight per household imposes an additional constraint on the estimation procedure, one might expect variance to suffer somewhat, particularly if no additional auxiliary information is brought to bear in estimation.

Variances for the post-stratification estimator were estimated using the Keyfitz method (1957) with PSU's (primary sampling units) or collapsed PSU's as replicates. The least-squares variances were estimated using the method described in Fuller (1975). To ensure comparability, variances for several characteristics estimated by means of post-stratification were calculated using the Fuller technique and compared to those from the Keyfitz approach. In all cases the two sets of variance estimates were very close (within one or two percent).

Table 3 summarizes the estimated efficiencies of the least-squares estimators relative to post-stratification for the characteristics considered in Table 1. The efficiency gains for estimates of economic families are substantial. Estimates of persons employed and of unattached persons also appear to gain somewhat; however, the variance reductions for these characteristics are small, with the exception of employed in the Atlantic Region, particularly when children are included in the weighting. Interestingly average family sizes in the Atlantic Region are higher than in the rest of the country, although it is not clear how this would affect estimates of employed persons. The variances for the characteristic unemployed are essentially unaffected by the least-squares procedure. One can probably expect these results to hold in general, i.e. for arbitrary characteristics. Although the one-weight-per-household criterion is a restrictive one for estimates of characteristics of persons, the least-squares estimators appear to compensate through the additional "explanatory" variables of the linear model, i.e. the household means of all auxiliary variables. The above preliminary results suggest that individual and family estimation could be integrated at little or no loss in efficiency for estimates of persons.

5. PLANS FOR FURTHER STUDY

The results presented in this paper are preliminary, and a more extensive empirical evaluation of the properties of the least-squares estimator is currently under way, with particular attention being given to the behaviour of estimates over time and to efficiencies for a larger group of characteristics relative to estimates produced with the Labour Force Survey's current raking ratio estimator. The foregoing results have suggested that at least for some characteristics of persons, the "explanatory power" of the age-sex composition of a household is at least as great as that of the age-sex group alone. It will be instructive to see if the relative efficiencies will be as favourable for characteristics more strongly correlated with age-sex. In addition although in practice negative weights have been uncommon, it is likely that some procedure must be developed to deal with them when they occur. Among the possibilities one might consider would be to accord them outlier treatment or perhaps to forestall their occurrence by imposing some bound on changes to the weights (Zieschang 1987). Finally it would be useful to make explicit the undercoverage model underlying the least-squares estimator to permit an evaluation of the model on its own merits.

ACKNOWLEDGEMENTS

The author would like to thank F. Scheuren for his comments and suggestions regarding this paper.

REFERENCES

- ALEXANDER, C.H. (1987). A class of methods for using person controls in household weighting. *Survey Methodology*, 13, 183-198.
- AUGER, E. (1987). Family data from the Canadian personal income tax file. In *Statistics of Income and Related Administrative Record Research: 1986-1987*, (Eds. W. Alvey and B. Kilss), Washington, D.C.: Internal Revenue Service, 177-184.
- BETHLEHEM, J.C., and KELLER, W.A. (1987). Linear weighting of sample survey data. *Journal of Official Statistics*, 3, 141-153.
- FULLER, W.A. (1975). Regression analysis for sample surveys. *Sankhyā*, C37, 117-132.
- GOSSELIN, J.-F., and THÉROUX, G. (1980). 1976 Census of Canada Quality of Data Series I: Sources of Error - Coverage. Catalogue No. 99-840, Statistics Canada.
- KEYFITZ, N. (1957). Estimates of sampling variance where two units are selected from each stratum. *Journal of the American Statistical Association*, 52, 503-510.
- OH, H.L., and SCHEUREN, F. (1978). Multivariate raking ratio estimation in the 1973 exact match study. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 716-722.
- PAUL, E.C., and LAWES, M. (1982). Characteristics of respondent and non-respondent households in the Canadian Labour Force Survey. *Survey Methodology*, 8, 48-85.
- PLATEK, R., and SINGH, M.P. (1976). Methodology of the Canadian Labour Force Survey. Catalogue No. 71-526, Statistics Canada.
- RUBIN, D.B. (1976). Inference on missing data. *Biometrika*, 63, 581-592.
- SINGH, M.P., DREW, J.D., and CHOUDHRY, G.H. (1984). Post '81 censal redesign of the Canadian Labour Force Survey. *Survey Methodology*, 10, 127-140.
- SCHEUREN, F., OH, H.L., VOGEL, L., and YUSKAVAGE, R. (1981). Studies from Interagency Data Linkages, Report No. 10: Methods of Estimation for the 1973 Exact Match Study. U.S. Department of Health and Human Services, Social Security Administration, SSA Publication No. 13-11750.
- WRIGHT, R.L. (1983). Finite population sampling with multivariate auxiliary information. *Journal of the American Statistical Association*, 78, 879-884.
- ZIESCHANG, K.D. (1986). A generalized least squares weighting system for the Consumer Expenditure Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.
- ZIESCHANG, K.D. (1987). Sample weighting methods and estimation of totals in the Consumer Expenditure Survey. Unpublished manuscript, U.S. Bureau of Labor Statistics.

Modified Raking Ratio Estimation

H. LOCK OH and FRITZ SCHEUREN¹

ABSTRACT

A hybrid technique is described that employs both conventional and raking ratio estimation to handle the case when the population frequencies N_{ij} in a two-dimensional table are known, but some of the observed frequencies n_{ij} are small (or zero). Results are provided on the approach taken as it has evolved in the Corporate Statistics of Income Program over the last several years. Changes are still being considered and these will be discussed as well.

KEY WORDS: Raking ratio estimation; Conventional ratio estimation; Conditional bias and variance.

1. INTRODUCTION

Raking ratio estimation, or simply "raking," is a widely used technique in sample surveys. Applications differ depending on the nature of the sample design, the extent of the auxiliary information available and the presence of various nonsampling errors (such as might arise because of nonresponse or undercoverage).

Raking was first proposed by Deming and Stephan (1940) as a way of assuring consistency between complete count and sample data from the 1940 U.S. Census of Population. The originators themselves elaborated their ideas early on (Deming 1943; Stephan 1942). Since then, perhaps because of the basic intuitive appeal of the iterative algorithm employed, there have been several wholly independent rediscoveries of the technique (Fienberg 1970).

Advances and modifications have also been numerous. For example, important theoretical work on convergence of the algorithm was done by Ireland and Kullback (1968). As might be expected, practitioners at Statistics Canada, and also at the U.S. Bureau of the Census, have deeply studied the application of raking in census and survey taking, especially in situations where the raking is not allowed to proceed to complete convergence (e.g., Brackstone and Rao 1979; Fan *et al.* 1981). A reasonably complete bibliography of the statistical research on raking prior to 1978 can be found in Oh and Scheuren (1978b).

In many treatments of raking, it is assumed that two (or more) sets of marginal population totals, say $N_{i\cdot}$ and $N_{\cdot j}$, are known, but that the interior of the table N_{ij} can only be estimated from the sample. When the N_{ij} are also known, the usual ratio estimator with weights N_{ij}/n_{ij} would be the natural choice, unless the corresponding sample sizes n_{ij} are "too small."

The present paper describes a hybrid technique that employs both conventional and raking ratio estimations to handle the case when the population cell frequencies N_{ij} are known, but some of the observed frequencies n_{ij} are small (or zero). In Section 2, we describe our approach. Some empirical results from the application of the method to our Corporate Statistics of Income Program are covered in Section 3. In Section 4, we conclude with a brief summary and some plans for the future.

H. Lock Oh and Fritz Scheuren, Statistics of Income Division, Internal Revenue Service, 1111 Constitution Avenue N.W., Washington, D.C. 20224, U.S.A.

2. RAKING RATIO ESTIMATION

2.1 General Considerations

Raking ratio estimation usually assumes that two (or more) marginal population totals, say, $N_{i.}$ and $N_{.j}$ are known, but that the interior of the table N_{ij} can only be estimated from the sample by, say, \tilde{N}_{ij} , where graphically (Deming 1943) we have

	1	2	...	S	
1	N_{11}	N_{12}	...	N_{1S}	$N_{1.}$
2	N_{21}	N_{22}	...	N_{2S}	$N_{2.}$
...
i	N_{ij}	...	$N_{i.}$
...
R	N_{R1}	N_{R2}	...	N_{RS}	$N_{R.}$
	$N_{.1}$	$N_{.2}$...	$N_{.S}$	N

with $i = 1, \dots, R$ and $j = 1, \dots, S$. The corresponding sample count table is

	1	2	...	S	
1	n_{11}	n_{12}	...	n_{1S}	$n_{1.}$
2	n_{21}	n_{22}	...	n_{2S}	$n_{2.}$
...
i	n_{ij}	...	$n_{i.}$
...
R	n_{R1}	n_{R2}	...	n_{RS}	$n_{R.}$
	$n_{.1}$	$n_{.2}$...	$n_{.S}$	n

In simple random sampling, the raking algorithm begins by setting

$$\tilde{N}_{ij} = \frac{N}{n} n_{ij}, \quad (2.1)$$

and then proceeds by proportionately scaling the \tilde{N}_{ij} such that the relations

$$\sum_j^S \tilde{N}_{ij} = N_{i.} \quad (2.2)$$

and

$$\sum_i^R \tilde{N}_{ij} = N_{.j} \quad (2.3)$$

are satisfied in turn. Each step in the algorithm begins with the results of the previous step, with the \tilde{N}_{ij} continuing to change; the process terminates either after a fixed number of steps or when expressions (2.2) and (2.3) are simultaneously satisfied to the closeness desired. (See Oh and Scheuren (1983) for further details; see Ireland and Scheuren (1975) for generalizations to multi-way tables and the handling of computational efficiency issues.)

By an application of the theory of minimum discrimination information (Kullback 1968), it can be shown (e.g., Ireland and Kullback 1968) that, under some regularity conditions if only the $N_{i.}$ and $N_{.j}$ are known, the \tilde{N}_{ij} obtained by raking to convergence are asymptotically unbiased, normally distributed and minimum variance (i.e., best asymptotically normal, or BAN, estimators). Theoretical results of this kind are partly what motivates the raking estimator for a general survey characteristic Y_{ijk} (e.g., income or assets), where we are interested in estimating the population total

$$Y = \sum_i^R \sum_j^S \sum_k^{N_{ij}} Y_{ijk} \tag{2.4}$$

with, say, the statistic

$$\tilde{Y} = \sum_i^R \sum_j^S \frac{\tilde{N}_{ij}}{n_{ij}} \left(\sum_k^{n_{ij}} Y_{ijk} \right). \tag{2.5}$$

Typically, of course, in survey processing a raking weight

$$\tilde{W}_{ij} = \frac{\tilde{N}_{ij}}{n_{ij}} \tag{2.6}$$

is placed on each individual record on the file for ease of handling. It is important to note that a feature of the raking algorithm is that if $n_{ij} = 0$ then necessarily $\tilde{N}_{ij} = 0$. For convenience, let $\tilde{W}_{ij} = 0$ in such cases as well.

Our interest below will be mainly on the conditional properties of the various estimators being examined. Such an approach has considerable appeal, as advocated by Holt and Smith (1979) and Rao (1985). (As an aside, it may be worth noting that Brackstone and Rao (1979), among others, have looked at the conditional behavior of the raking estimator. They conditioned, however, on the sample marginals $n_{i.}$ and $n_{.j}$.)

2.2 Conditional Bias

Following Oh and Scheuren (1983) we focus primarily in this paper on the conditional properties of \tilde{Y} , given $\underline{n} = (n_{11}, n_{12} \dots, n_{RS})$. In particular, let \bar{Y}_{ij} be the population mean for the ij -th subgroup. Then the conditional expected value of \tilde{Y} is

$$E(\tilde{Y} | \underline{n}) = \sum_i^R \sum_j^S \tilde{N}_{ij} \bar{Y}_{ij} = Y + \sum_i^R \sum_j^S (\bar{Y}_{ij} - \bar{Y}) (\tilde{N}_{ij} - N_{ij}). \tag{2.7}$$

Thus \tilde{Y} is conditionally biased with the importance of the bias depending on the structure of the population and whether or not the raking is to convergence. (Of course, when raking to convergence, unconditionally $E(\tilde{N}_{ij}) = N_{ij}$ asymptotically.)

Employing the usual analysis of variance conventions (e.g., Scheffé 1959)

$$(\bar{Y}_{ij} - \bar{Y}) = (\bar{Y}_i - \bar{Y}) + (\bar{Y}_j - \bar{Y}) + (\bar{Y}_{ij} - \bar{Y}_i - \bar{Y}_j + \bar{Y}); \quad (2.1)$$

hence the conditional bias, given \underline{n} , is expressible as

$$\begin{aligned} \text{Bias}(\tilde{Y} | \underline{n}) &= \sum_i^R (\bar{Y}_i - \bar{Y}) (\tilde{N}_i - N_i) + \sum_j^S (\bar{Y}_j - \bar{Y}) (\tilde{N}_j - N_j) \\ &+ \sum_i^R \sum_j^S (\bar{Y}_{ij} - \bar{Y}_i - \bar{Y}_j + \bar{Y}) (\tilde{N}_{ij} - N_{ij}). \end{aligned} \quad (2.2)$$

If the raking is to convergence, then the first two terms of the conditional bias become zero. For the third term of the conditional bias to be zero for either form of raking, it is sufficient that the Y_{ij} be such that there is no interaction. In large-scale surveying with many variables this is unrealistic to assume; nonetheless, in practice the interaction is often a minor part of the decomposition of Y_{ij} ; consequently, the raking ratio estimator may, in many cases, have small biases even in moderate sample sizes.

2.3 Conditional Variance

Conditional and unconditional approaches to the variance of the raking ratio estimator have been extensively examined (e.g., Binder 1983; Causey 1972; Bankier 1986; Fan *et al.* 1981; Brackstone and Rao 1979). In our own early work (described in Section 3.2), we have employed replication techniques (e.g., Leszcz, Oh and Scheuren 1983). The replication methods used (which were equivalent to conditioning on the sample marginals) proved expensive, unwieldy, and somewhat unstable, leading us to a simpler attack on the conditional variance estimation problem (albeit the level of conditioning was deeper).

To motivate the approach we are currently taking, consider the conditional variance of \tilde{Y} , given \underline{n} . Now it can be shown by a slight extension of Oh and Scheuren (1983) that

$$\text{Var}(\tilde{Y} | \underline{n}) = \sum_i^R \sum_j^S n_{ij} (\tilde{W}_{ij})^2 \left(1 - \frac{n_{ij}}{N_{ij}}\right) V_{ij} \quad (2.10)$$

where the V_{ij} are the population variances of the ij -th subgroup and if $N_{ij} = 0$ or 1 we define $V_{ij} = 0$. (We are also employing the convention in expression (2.10) that $0/0 = 0$.)

Expression (2.10) holds whether or not the raking goes to convergence. Despite this it has been little studied because it cannot be readily adapted to estimate the conditional variance. The principal difficulty, of course, lies in our inability to calculate stable estimators of the V_{ij} when the n_{ij} are small. To overcome this problem we began looking at collapsing techniques based on the size of the raking weight. First, we let \tilde{W}_{ij} approximate N_{ij}/n_{ij} which gives us

$$\text{Var}(\tilde{Y} | \underline{n}) \approx \sum_i^R \sum_j^S n_{ij} \tilde{W}_{ij} (\tilde{W}_{ij} - 1) V_{ij}. \quad (2.11)$$

Now if the \tilde{W}_{ij} are ordered from smallest to largest and if they vary over a narrow range then averaging them into (ordered) groups of, say, about $n_g \geq 25$ observations each will

alter the value of expression (2.11) very little. It will, however, allow us to calculate collapsed post-stratum variance estimates for the V_{ij} . This is the approach we have taken in Section 3.

One final point should be noted. The alternative proposed here is stable and fairly easy to calculate. Our limited empirical work, however, is inconclusive on the method's utility and, while we feel the method is worthy of discussion, we are in no sense advocating its general use at this time.

3.4 Modified Raking Estimation

As we have noted, under fairly general conditions the \tilde{N}_{ij} are BAN estimators. This does not mean, however, that \tilde{Y} will share all these properties. Indeed, if the variables used in the raking are not highly correlated with the characteristic Y , the estimator \tilde{Y} may suffer some degradation in variance relative, say, to a simple ratio estimator

$$\tilde{Y} = \sum_j^S \left(\frac{N_j}{n_j} \right) \left(\sum_i^R \sum_k^{n_{ij}} Y_{ijk} \right). \quad (2.12)$$

Typically, of course, experience has shown that both positive and negative impacts may occur in the same sample. The practitioner's problem is somehow to keep the positive effects while minimizing the negative ones.

There seems to be no general solution to this dilemma but we have had some limited successes, in our application settings, with two techniques that may be of wider interest (see subsections 3.2 and 3.3 for results).

In most treatments of raking, it is assumed that the marginal population totals $N_{i.}$ and $N_{.j}$ are known; and that the interior of the table N_{ij} can only be estimated from the sample. In our setting we actually have the population values N_{ij} and are employing raking as a way of systematically handling cells in the table where the n_{ij} are small. Conventional collapsing alternatives exist here, of course (e.g., Cochran (1977) Fuller (1966)); but seemed unsuitable for reasons that will be explained later.

It may be possible to agree that raking is a satisfactory way of handling the small cells in this setting; but what about the larger ones? Surely it would be better to use the conventional simple ratio estimator in the large cells. Indeed, if this were done, the conditional bias for these "large" cells would be zero; but what would be the effect on the rest of the cells? This line of reasoning suggested that we employ a hybrid estimation method where, for cells where the n_{ij} was large, the conventional simple ratio estimator is used. These cells are then removed from the population and sample tables, and the remaining sample cells are raked to the adjusted population marginals.

For the remaining smaller cells, a second procedure was introduced to reduce the possible negative impacts of the raking on certain variables. We bounded the raking so that the weights \tilde{W}_{ij} did not vary "too much" from the initial weight. (This kind of constraint is often employed, by the way, in simple ratio estimation, e.g., Hanson 1978.)

The approach to bounded raking ratio estimation is similar to that when "large" sample counts are available in a single cell. That is, it is similar in that, for the cell that is to be constrained, we bound the \tilde{W}_{ij} ; then take the estimated population total $\tilde{N}_{ij} = \tilde{W}_{ij} n_{ij}$ for that cell and the sample n_{ij} for that cell out of the population and out of the sample tables respectively); and then adjust the remaining observations.

Three problems exist with these partial "solutions." First there is the (uncomfortable) arbitrariness of the definitions of a "large" cell, and of a weighting factor that varies "too much" from its initial value. A related concern was why, if we were willing to use simple ratio estimation for "large" cells, conventional collapsed stratum techniques could not be

used for the remaining cells. The third problem has to do with the properties of the raking algorithm's convergence when we employ this hybrid. It is quite clear, for example, from the research that has been done on raking that tables with too many zeros in them will be very unstable and the raking may not converge (e.g., Oh and Scheuren 1978a and 1978b; Ireland and Scheuren 1975). This is of particular concern since the effect of both our modifications is to introduce zeros into the table. If these zeros are strategically placed, or better *misplaced*, then this could have a very serious detrimental impact on the rate of convergence and, even, on the quality of the estimators. Our recommendation before starting was, therefore, that the number of times that these procedures were employed would have to be fairly small. It is beyond the scope of the present paper to resolve these concerns in general (if indeed that is possible). In Section 3, however, we will consider them further for the applied setting in which we did this work, and also will return to them in Section 4, when discussing areas for future study.

3. RAKING IN THE CORPORATE STATISTICS OF INCOME PROGRAM

3.1. Background

The U.S. Internal Revenue Service has produced statistics from corporate tax returns annually for over 70 years. Corporate data are, in fact, a mainstay of the so-called Statistics of Income Program, which is the name collectively given to all of the non-administrative statistical series produced by the Internal Revenue Service for public consumption.

Until 1951, corporate statistics were based on a complete census of the returns filed. Since then, a stratified probability sample has been employed, currently running in size at about 90,000 returns annually (from about 3,000,000 returns filed). Assets and income are the principal stratifying variables (Jones and McMahon 1984). Stratification by industry has long been considered, as well, but the quality of the industry coding as self-reported by taxpayers seemed insufficient to justify this step on a wholesale basis. Typically, for example, at the minor industry level perhaps 20 percent or more of the self-reported codes are changed during statistical processing. Nonetheless, because of the importance of industry statistics, efforts to use administrative data by industry to post-stratify the sample still seemed warranted and have been pursued over many years (e.g., Westat, Inc. 1974; Leszcz, Oh, and Scheuren 1983).

In a pilot post-stratification study done by Westat during the early 1970's, substantial improvements in standard errors were achieved for a number of variables, notably Total Receipts (where a reduction of about 12 percent occurred). Some increases in standard errors took place, however, for variables not closely related to industry (e.g., distribution to shareholders), but these were minor. To handle small cells, Westat used conventional collapsed stratum techniques to combine industry post-strata within the then-existing sample strata. Concerns continued to exist about the quality of the administrative industry data especially for small cells; in any case, due to other operational priorities, the Westat approach was never implemented.

A major series of budget cuts occurred during the 1980-1982 period, and these forced a number of changes in the sample designs and estimation procedures across nearly all the studies that make up the Statistics of Income Program (e.g., Hinkins and Scheuren 1986; Scheuren, Schwartz, and Kilss 1984); in particular, the corporate study experienced sample size cuts during this period which, although later partially rescinded, reopened the issue of post-stratification by industry.

A raking ratio estimation approach to post-stratification seemed to have appeal over what Westat had done. One of the reasons for this was that concerns about the quality of the marginal administrative totals, by industry, were not as great as for the individual cells. The work of implementing a collapsing scheme could be completely avoided, as well.

3.2 Early Modified Raking Results

When we implemented a pure raking scheme for the Tax Year 1979 sample, our principal customers expressed concerns about what we had done. They were particularly worried about the potential for large adjustment factors having an adverse effect on certain statistics. We, in turn, having seen the results ourselves, were concerned that we had not done an adequate job for those industry-sample stratum combinations where the number of sample observations were large. As a consequence, these results were never used and the 1979 Tax Year statistics were published employing normal stratified sampling estimation (NORM).

Research continued, however, and in 1983, a paper was given comparing the root mean square errors of six different variations of raking both with each other and with what we had been doing previously (Leszcz, Oh, and Scheuren 1983). Three "pure" raking alternatives were looked at:

PRRE: "Classical" raking ratio estimation to convergence (Deming and Stephan 1940);

PRRE (200): Simple ratio adjustment of cells with samples of 200 returns or more and "classical" raking of the remaining cells to convergence; and

PRRE (400): Simple ratio adjustment of cells with samples of 400 returns or more and "classical" raking of the remaining cells to convergence.

In addition, three versions of bounded raking ratio estimation were examined, all with the bounds set at $(\sqrt{2/3}, \sqrt{3/2})$. These were:

BRRE: Bounded raking ratio estimation (2 cycles);

BRRE (200): Simple ratio adjustment of cells with samples of 200 and bounded raking (2 cycles) of the remaining cells; and

BRRE (400): Simple ratio adjustment of cells with samples of 400 and bounded raking (2 cycles) of the remaining cells.

For the bounded raking we were initially not sure that complete convergence was possible; hence, we made an operational simplification and only cycled through the constraint equations, e.g., (2.2) and (2.3), twice.

To make the root mean square error (RMSE) comparison, pseudo-replicate half-samples were drawn, each designed in the same way as the overall sample. The procedure involved: (1) construction of the half-samples; (2) two-way classification – by original sample stratum and major industry (post-stratum) – of sample counts for each half-sample; (3) derivation of a set of weights for each half-sample for each estimator; (4) calculation of estimates of selected items by applying the weight to sample values for each half-sample; and (5) calculation of the RMSE, based on the variations in the estimates that each half-sample produced. For cost reasons only 14 sets of half samples were used.

The resultant summary tabulation presented as Table I reveals what one would have expected of the number of returns. Near 100 percent reductions occurred for the PRRE, PRRE(200), and PRRE(400) estimates. Application of the bounding limits $\sqrt{2/3}$ and $\sqrt{3/2}$, and not cycling to convergence, decreased the magnitude of these reductions; however, they were still substantial. As Table I also indicates, for Total Receipts, a key variable, there were also improvements, although much less sizable.

Table 1
Reduction in Root Mean Square Error (RMSE)
as a Percent of Corresponding Normal Stratified Sampling RMSE

Estimator	Number of Returns	Total Receipts	Jobs Credit
"Pure" raking ratio estimators:			
PRRE	98.6	8.3	-3.0
PRRE (400)	98.6	9.2	-3.0
PRRE (200)	98.6	11.9	-3.0
Bounded raking ratio estimators:			
BRRE	74.0	13.8	+1.0
BRRE (400)	73.4	15.6	+1.0
BRRE (200)	72.3	17.4	+1.0

Note: The percentages shown are simple averages of the percent reductions in each of the 56 major industry groups used in the post-stratification. Notice that the percentage improvements for the "number of returns" column are nearly but *not* 100 percent for the PRRE estimators. This occurs because the raking took place for all corporations with both the N_{ij} and n_{ij} defined on this basis; however, only active corporations (about 90 percent) were tabulated. The BRRE estimators in the "number of returns" column differ from each other and from the PRRE estimates because the cycling was not to convergence. This has subsequently been changed, beginning with Tax Year 1980.

Jobs Credit results in Table 1 are included to illustrate the expected tradeoff that can exist for items not closely related to industry. In particular, we see that in some cases there are (modest) increases in the root mean square errors for this item, due presumably to the fact that this field is less dependent upon the industry groupings utilized in this research.

It should be noted that, for Total Receipts, the decreases shown in the root mean square error, from the initial (NORM) estimate to that utilizing raking ratio estimation, all compare favorably with the Westat pilot study results. While we are encouraged by this comparison, a great deal has changed over the decade between the earlier Westat results and those in Leszcz, Oh and Scheuren (1983). What would really be telling, and what has not been done, is to compare conventional collapsing schemes with our modified approach to raking *on the same data set*.

One final point about Table 1; it reflects improvements in RMSE when tabulating by the administrative industry information which was used in the post-stratification. Because of differences between the administratively and statistically assigned classifications by industry the figures shown in this table are therefore likely to overstate the improvements being achieved in our published statistics, since so many entities (over 20 percent) are recoded during the in-depth processing done of our corporate sample.

3.3 Current Modified Raking Results

Beginning with Tax Year 1980, we began to regularly produce and publish our corporate statistics using the bounded raking ratio estimator BRRE(200) (U.S. Department of Treasury 1984). For Tax Years 1983 and later, we made the modifications described in Section 2. so that approximate conditional variances could be calculated. These were first published for Tax Year 1984 (U.S. Department of Treasury 1987). Also, in an effort to confirm the earlier results, we undertook for Tax Year 1984 to compare the conditional variance of the modified raking method being employed with the variance that would have been estimated had we used normal stratified sampling estimation. Before discussing the limited comparison made, it might be worthwhile giving some of the application details on the corporate setting for 1984.

In our earlier work (Leszcz, Oh and Scheuren, 1983), and for 1984, the entire corporate return population of IRS Forms 1120 and 1120S was tallied into 58 major industry groups. For 1984, industry was cross-classified by 14 sample strata in each of the two processing years during which the sample had to be selected. Some of the major industries were so sparse that we immediately collapsed the industry detail to 56 groups. This still left a very large table (of 1568 cells).

It may be of interest to note that there were 414 "natural" zero cells in the population and an additional 125 zero cells arising in the sample. Before raking we removed 96 cells that had 200 or more sample observations; these cells were then each ratio adjusted separately. (In all, 57 percent of the Forms 1120 and 1120S corporate sample were so adjusted.) Finally, there were 73 cells that had to be bounded during the raking itself. This meant that altogether in the raking step there were 708 or 45 percent of the cells being treated as zeroes.

The raking was initiated by introducing the normal stratified estimator into each cell of the table. The marginal constraints imposed were (1) by industry and sampling period, and (2) by sample strata and sampling period. In the published statistics for 1984, and in the comparisons made here, the raking did not go to convergence; it was just carried out for two cycles. (Incidentally, concerns about the conditional bias of this approach have led us to rake our 1985 sample data to convergence.)

The results of the efforts for 1984 were to reduce the overall and industry-by-industry standard errors for frequencies by substantial amounts – only about half as much, however, as is shown in Table 1. Similar dampened improvements occurred for Total Receipts (8.7 percent) with many variables like Jobs Credit and Net Income experiencing little or no change in their standard errors overall (see U.S. Department of Treasury 1987, for details). As already noted, conditioning may be part of the reason for this difference (Holt and Smith 1979). The original results were conditional on the sample marginals n_i and n_j ; the later figures employed a deeper level of conditioning.

We are still examining other possibilities as to why the improvements are more modest than we found in the earlier work. Some obvious possibilities are the way we grouped the data from the smaller cells, including the consequent averaging of the weighting factors \tilde{W}_{ij} , and the collapsed variance estimation of the V_{ij} . Tabulating the data using our statistical industry coding, rather than the administrative coding, as in Table 1, may have been a major factor.

4. CONCLUSIONS AND AREAS FOR FURTHER STUDY

4.1 General

The modified raking approach for our corporate sample certainly seems to be an improvement over the normal stratified sampling approach taken formerly. There are, however, a number of unsettling *ad hoc* aspects of the method that trouble us. For instance, the connection between conventional collapsed stratum techniques and our modified raking procedure needs more study. Exploring changes in estimation techniques is not enough, however. More work on the basic sample design appears needed too. Finally, the variance approximation being used needs further looking at. We may well have paid a high price for stability and ease of calculation. As noted earlier, the statistical literature is full of good alternatives, and these deserve to be examined in a full-scale comparison with what we are currently doing.

4.2 Estimation Issues

There is considerable intuitive appeal in developing a post-stratification method that *smoothly* increases the degree of conditioning from just using marginal totals to using some

or all of the interior population counts as well. Our current approach has an embarrassing *ad hoc* flavor. Frankly, we see it just as a stop gap until we can increase the quality of the underlying administrative data by industry. Our main concern is to reduce response variation arising from taxpayer or processing errors. Even if we are unsuccessful in improving the administrative data directly, it may be possible to dampen the response error effects by looking at the tables by industry and sample stratum over several years. This is planned and may allow us to integrate, in a more complete way, raking on the one hand and collapsed post-stratum estimation on the other.

4.3 Design Issues

Improved administrative data by industry has obvious uses at the design stage. At the present time, coefficients of variation differ quite widely by industry, with the smaller industries being very poorly represented. No amount of after-the-fact post-stratification can correct for this completely. Improving the balance by industry, and over time, appear to be top priorities (e.g., Hinkins, Jones and Scheuren 1987).

ACKNOWLEDGMENTS

As is true of nearly all applied work, the authors have many people to thank for the results in this paper. Homer Jones, who has responsibility for the corporate sample, made important contributions, ably assisted by Richard Collins. Nat Shaifer provided illuminating statistics on the degree of comparability of the administrative and statistical coding of industry. Mike Leszcz played the leading role in presenting the earlier results on this topic at the 1983 American Statistical Association meetings in Toronto. The paper profited from conversations with Rod Little and many helpful suggestions were made by the referees. Bettye Jamerson assisted in the manuscript preparation.

REFERENCES

- BANKIER, M.D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association*, 81, 1074-1079.
- BINDER, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.
- BRACKSTONE, G.J., and RAO, J.N.K. (1979). An investigation of raking ratio estimators. *Sankhyā* Ser. C, 41, 97-114.
- CAUSEY, B.D. (1972). Sensitivity of raked contingency table totals to changes in problem conditions. *Annals of Mathematical Statistics*, 43, 656-658.
- COCHRAN, W.G. (1977). *Sampling Techniques*, (3rd ed.). New York: John Wiley.
- DEMING, W.E. (1943). *Statistical Adjustment of Data*. New York: Dover.
- DEMING, W.E., and STEPHAN, F.F. (1940). On a least square adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11, 427-444.
- FAN, M.C., WOLTMAN, H.F., MISKURA, S.M., and THOMPSON, J.H. (1981). 1980 Census: variance estimation procedure. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 176-181.
- FIENBERG, S.E. (1970). An iterative procedure for estimation in contingency tables. *Annals of Mathematical Statistics*, 907-917.

- FULLER, W.A. (1966). Estimation employing post strata. *Journal of the American Statistical Association*, 61, 1172-1183.
- HANSON, R. (1978). The Current Population Survey: design and methodology. Technical Paper No. 40, U. S. Bureau of the Census.
- HINKINS, S., JONES, H., and SCHEUREN, F. (1987). Updating tax return selection probabilities in the corporate Statistics of Income program. A paper presented at the International Symposium on Statistical Uses of Administrative Data, Ottawa, Canada, November 23-25, 1987.
- HINKINS, S., and SCHEUREN, F. (1986). Hot deck imputation procedure applied to a double sampling design. *Survey Methodology*, 12, 181-195.
- HOLT, D., and SMITH, T. M. F. (1979). Post-stratification. *Journal of the Royal Statistical Society, Series A*, 142, 33-46.
- IRELAND, C.T., and KULLBACK, S. (1968). Contingency tables with given marginals. *Biometrika*, 55, 179-188.
- IRELAND, C.T., and SCHEUREN, F.J. (1975). The rake's progress. *Computer Programs for Contingency Table Analysis*, Washington, DC: The George Washington University, 155-216.
- JONES, H., and MCMAHON, P.B. (1984). Sampling corporation income tax returns for Statistics of Income, 1951 to present. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 269-274.
- KULLBACK, S. (1968). *Information Theory and Statistics*. New York: Dover.
- LESZCZ, M., OH, H.L., and SCHEUREN, F. (1983). Modified raking estimation in the corporate SOI program. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 107-111.
- OH, H.L., and SCHEUREN, F.J. (1978a). Multivariate raking ratio estimation in the 1973 Exact Match Study. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 716-722.
- OH, H.L., and SCHEUREN, F.J. (1978b). Some unresolved application issues in raking ratio estimation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 723-728.
- OH, H.L., and SCHEUREN, F.J. (1983). Weighting adjustments for unit nonresponse. In *Incomplete Data in Sample Surveys, Volume 2, Theory and Bibliographies*, (Eds, W.G. Madow, I. Olkin and D.B. Rubin), New York: Academic Press, 143-184.
- RAO, J.N.K. (1985). Conditional inference in survey sampling. *Survey Methodology*, 11, 15-31.
- SCHEFFÉ, H. (1959). *The Analysis of Variance*. New York: John Wiley, 106-119.
- SCHEUREN, F., SCHWARTZ, O., and KILSS, B. (1984). Statistics from individual income tax returns: quality issues and budget cut impact. *Review of Public Data Use*, 12, 55-67.
- STEPHAN, F.F. (1942). Iterative method of adjustment sample frequency tables when expected margins are known. *Annals of Mathematical Statistics*, 13, 166-178.
- U. S. DEPARTMENT OF TREASURY (1984). *Statistics of Income - 1980 Corporation Income Tax Returns*, Publication 16, Internal Revenue Service, U.S. Department of Treasury.
- U. S. DEPARTMENT OF TREASURY (1987). *Statistics of Income - 1984 Corporation Income Tax Returns*, Publication 16, Internal Revenue Service, U.S. Department of Treasury.
- WESTAT, INC. (1974). Results of a study to improve sampling efficiency of statistics of corporation income. Bethesda, Maryland (unpublished).

Comparison of the Horvitz-Thompson Strategy with the Hansen-Hurwitz Strategy

S.G. PRABHU-AJGAONKAR¹

ABSTRACT

The Hansen-Hurwitz (1943) strategy is known to be inferior to the Horvitz-Thompson (1952) strategy associated with a number of IPPS (inclusion probability proportional to size) sampling procedures. The present paper presents a simpler proof of these results and therefore has some pedagogic interest.

KEY WORDS: Sampling strategies; Inclusion probability proportional to size; Positive definite quadratic form.

1. INTRODUCTION

Let U be a finite population consisting of N identifiable units $[U_1, U_2, \dots, U_N]$. With the i -th unit of the population U_i are associated two numbers X_i and Y_i , where X_i 's are known and Y_i 's are fixed but unknown. Generally, X_i represents a measure of size of U_i which is highly correlated with Y_i .

For estimating the population total $T_y = Y_1 + Y_2 + \dots + Y_N$, the Hansen and Hurwitz (1943) strategy consists of selecting with replacement n population units with probability proportional to X_i , and using the unbiased estimator

$$t_{HH} = \frac{1}{n} \sum_{r=1}^n \frac{y_r}{p_r}$$

where $p_r = X_r/T_x$, $T_x = X_1 + X_2 + \dots + X_N$, and y_r ($r=1, 2, \dots, n$) represents the outcome at the r -th draw. It is easy to show, noting that $\sum Z_i = 0$,

$$\text{Var}(t_{HH}) = \sum_{i=1}^N \frac{Z_i^2}{np_i} \quad (1)$$

where $Z_i = Y_i - p_i T_y$, $i=1, 2, \dots, N$.

When population units are selected without replacement, Horvitz and Thompson (1952) proposed the unbiased estimator

$$t_{HT} = \sum_{i=1}^n \frac{y_i}{\pi_i}$$

¹ S.G. Prabhu-Ajgaonkar, Department of Mathematics and Statistics, Marathwada University, Aurangabad 431004, India.

where π_i ($i=1, 2, \dots, N$) denotes the probability of including the i -th population unit U_i in the sample. Further, when π_i is proportional to X_i , the sampling procedure is termed an IPPS scheme. For such a sampling procedure,

$$\text{Var}(t_{HT}) = \sum_{i=1}^N \frac{Z_i^2}{np_i} + \sum_{i \neq j=1}^N Z_i Z_j \frac{\pi_{ij}}{n^2 p_i p_j} \quad (2)$$

where Z_i is given in (1), and π_{ij} ($i \neq j=1, 2, \dots, N$) represents the joint probability of including the i -th and j -th population units in the sample. When an IPPS procedure is specified π_{ij} can be further simplified.

From (1) and (2),

$$\phi = \text{Var}(t_{HT}) - \text{Var}(t_{HH}) = \sum_{i \neq j=1}^N Z_i Z_j \frac{\pi_{ij}}{n^2 p_i p_j}. \quad (3)$$

2. COMPARISON OF STRATEGIES

Midzuno (1952), Sen (1952) and Sankaranarayanan (1969) proposed IPPS sampling schemes for estimating T_y , using the Horvitz-Thompson estimator t_{HT} . The Midzuno-Sen scheme is feasible if

$$p_i = \frac{X_i}{T_x} > \frac{n-1}{n(N-1)}, i=1, \dots, N, \quad (4)$$

Sankaranarayanan's scheme requires the weaker condition

$$\sum_{j \in s} p_j > (n-1)/(N-1) \text{ for all } s \in S.$$

For both the schemes, the joint inclusion probabilities are given by

$$\pi_{ij} = \frac{n(n-1)}{N-2} \left(p_i + p_j - \frac{1}{N-1} \right).$$

Hence, from (3),

$$\phi = \frac{n(n-1)}{n^2(N-2)} \left[\sum_{i=1}^N \frac{Z_i^2}{p_i} \left(2 - \frac{1}{(N-1)p_i} \right) + \frac{1}{(N-1)} \left(\sum_{i=1}^N \frac{Z_i}{p_i} \right)^2 \right]. \quad (5)$$

The above expression is nonnegative if

$$P_i > \frac{1}{2(N-1)}, i=1, 2, \dots, N,$$

in which case the Horvitz-Thompson strategy is superior to the Hansen-Hurwitz strategy. The above restriction on X_i^* was first derived by Rao (1963) when $n=2$ and Midzuno-Sen scheme is employed, but it is interesting to note from (5) that the restriction remains the same even when n is greater than 2.

Chaudhuri (1975) and Mukhopadhyay (1975) independently derived the above for the Midzuno-Sen scheme.

Brewer (1963), Rao (1965) and Durbin (1967) proposed different IPPS schemes, for the case $n=2$, with the same inclusion probabilities,

$$\pi_{ij} = \frac{2p_i p_j}{1+k} \left(\frac{1}{1-2p_i} + \frac{1}{1-2p_j} \right) \text{ where } k = \sum_{i=1}^N \frac{p_i}{1-2p_i}.$$

These schemes are free from the restrictions on the p_i 's of the previous schemes. From (3),

$$\phi = \frac{1}{1+k} \sum_{i=1}^N \frac{Z_i^2}{1-2p_i} \geq 0,$$

so that the Hansen-Hurwitz strategy is again inferior to the Horvitz-Thompson strategy.

ACKNOWLEDGEMENTS

The author is indebted to the Editor, M.P. Singh, and a referee for their many helpful comments.

REFERENCES

- BREWER, K.R.W. (1963). A model of systematic sampling with unequal probabilities. *Australian Journal of Statistics*, 5, 5-13.
- CHAUDHURI, A. (1975). On some properties of the sampling scheme due to Midzuno. *Bulletin of Calcutta Statistical Association*, 23, 1-19.
- DURBIN, J. (1967). Design of multi-stage surveys for the estimation of sampling errors. *Applied Statistics*, 16, 152-164.
- HANSEN, M.H., and HURWITZ, W.N. (1943). On the theory of sampling from finite population. *Annals of Mathematical Statistics*, 14, 333-362.
- HORVITZ, D.G., and THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- MIDZUNO, H. (1952). On the sampling system with probability proportionate to sum of sizes. *Annals of Institute of Statistical Mathematics*, 3, 99-107.
- MUKHOPADHYAY, P. (1975). PPS sampling schemes to base HTE. *Bulletin of Calcutta Statistical Association*, 23, 21-44.
- RAO, J.N.K. (1963). On two systems of unequal probability sampling without replacement. *Annals of Institute of Statistical Mathematics*, 15, 67-72.
- RAO, J.N.K. (1965). On two simple schemes of unequal probability sampling without replacement. *Journal of the Indian Statistical Association*, 3, 173-180.

- SANKARANARAYANAN, K. (1969). An IPPS sampling scheme using Lahiri's method of selection. *Journal of the Indian Society of Agricultural Statistics*, 21, 58-66.
- SEN, A.R. (1952). Further developments of the theory and application of primary sampling units with special reference to the North Carolina agricultural population. Ph.D. Thesis, North Carolina State College, Raleigh.

ACKNOWLEDGEMENTS

The Survey Methodology Journal wishes to thank the following persons who have served as referees between January 1, 1987 and January 31, 1988. An asterisk indicates that the person served more than once.

- | | |
|----------------------------------------------------|----------------------------------------------------|
| D.W. Anderson, <i>National Institute of Health</i> | W.D. Kalsbeek, <i>University of North Carolina</i> |
| J. Armstrong, <i>Statistics Canada</i> | *G. Kalton, <i>University of Michigan</i> |
| H.R. Arora, <i>Transport Canada</i> | N.J. Kirkendall, <i>Alexandria, Virginia</i> |
| M. Bankier, <i>Statistics Canada</i> | G. Kriger, <i>Statistics Canada</i> |
| *K.G. Basavarajappa, <i>Statistics Canada</i> | H. Lee, <i>Statistics Canada</i> |
| G. Brackstone, <i>Statistics Canada</i> | J.M. Lepkowski, <i>University of Michigan</i> |
| D. Bellhouse, <i>University of Western Ontario</i> | *S. Kumar, <i>Statistics Canada</i> |
| L. Biggeri, <i>University of Florence</i> | R. Lachapelle, <i>Statistics Canada</i> |
| *D.A. Binder, <i>Statistics Canada</i> | E. Langlet, <i>Statistics Canada</i> |
| R.D. Burgess, <i>Statistics Canada</i> | S. Michaud, <i>Statistics Canada</i> |
| C.-M. Cassel, <i>Statistics Sweden</i> | D.G. Paton, <i>Statistics Canada</i> |
| N. Chinnappa, <i>Statistics Canada</i> | M. Podehl, <i>Statistics Canada</i> |
| *G.H. Choudhry, <i>Statistics Canada</i> | *J.N.K. Rao, <i>Carleton University</i> |
| D. Dodds, <i>Statistics Canada</i> | P.S.R.S. Rao, <i>University of Rochester</i> |
| D. Drew, <i>Statistics Canada</i> | G. Sande, <i>Statistics Canada</i> |
| M. Eagen, <i>Goss, Gilroy and Associates</i> | *I. Sande, <i>Statistics Canada</i> |
| J.L. Eltinge, <i>Iowa State University</i> | C.E. Särndal, <i>Université de Montréal</i> |
| I.P. Fellegi, <i>Statistics Canada</i> | *F. Scheuren, <i>U.S. Internal Revenue Service</i> |
| W.A. Fuller, <i>Iowa State University</i> | E.A. Schillmoeller, <i>Nielsen Media Research</i> |
| J.F. Gentleman, <i>Statistics Canada</i> | M. Sheridan, <i>Statistics Canada</i> |
| E. Gbur, <i>University of Arkansas</i> | A.C. Singh, <i>Statistics Canada</i> |
| *G.B. Gray, <i>Statistics Canada</i> | K.P. Srinath, <i>Statistics Canada</i> |
| M.A. Hidioglou, <i>Statistics Canada</i> | V. Tremblay, <i>Statplus</i> |
| D. Holt, <i>University of Southampton</i> | A. van Baaren, <i>Statistics Canada</i> |
| S. Ingram, <i>Statistics Canada</i> | *K.M. Wolter, <i>U.S. Bureau of the Census</i> |

Acknowledgements are also due to those who assisted during the production of the 1987 issues: B. Babcock (Text Editing), C. VanBastelaar (Photocomposition), G. Gaulin (Author Services) and M. Haight (Translation Services).

We would like to thank the staff of Social Survey Methods and Business Survey Methods Divisions who assisted in proofreading and verification. Finally we wish to acknowledge J. Clarke, E. Corriveau, J. Dufresne, M. Kent, C. Larabie, D. Lemire and N. Smalldridge for their support with coordination, typing and copy editing.

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 10, No. 2 and onward) of Survey Methodology as a guide and note particularly the following points:

Layout

- 1 Manuscripts should be typed on white bond paper of standard size ($8\frac{1}{2} \times 11$ inch), one side only, entirely double spaced with margins of at least $1\frac{1}{2}$ inches on all sides.
- 2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 4 Acknowledgements should appear at the end of the text.
- 5 Any appendix should be placed after the acknowledgements but before the list of references.

Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

Style

- 1 Avoid footnotes, abbreviations, and acronyms.
- 2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(\cdot)" and "log(\cdot)", etc.
- 3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 4 Write fractions in the text using a solidus.
- 5 Distinguish between ambiguous characters, (e.g., w, ω ; o, O, 0; l, 1).
- 6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

Figures and Tables

- 1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

References

- 1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
- 2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

avant de dactylographier votre texte pour le soumettre, priez d'examiner un numéro récent de Techniques d'enquête (à partir du vol. 10, n° 2) et de noter les points suivants:

Présentation

- 1. Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1 ½ pouce tout autour.
- 2. Les textes doivent être divisés en sections numérotées portant des titres appropriés. Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
- 3. Les remerciements doivent paraître à la fin du texte.
- 4. Toute annexe doit suivre les remerciements mais précéder la bibliographie.

Résumé

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

Rédaction

- 1. Éviter les notes au bas des pages, les abréviations et les sigles.
- 2. Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp() et log() etc.
- 3. Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
- 4. Écrire les fractions dans le texte à l'aide d'une barre oblique.
- 5. Distinguer clairement les caractères ambigus (comme w, ω; o, O; 0; 1, I).
- 6. Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.

Figures et tableaux

- 1. Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
- 2. Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois.)

Bibliographie

- 5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.
- 5.2 Exemple: Cochran (1977, p. 164).
La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

REMERCIEMENTS

La revue *Techniques d'enquête* désire remercier les personnes suivantes, qui ont accepté de faire la critique d'un article entre le premier janvier 1987 et le 31 janvier 1988. Une astérisque indique que la personne a participé plus d'une fois.

- D.W. Anderson, *National Institute of Health*
 J. Armstrong, *Statistique Canada*
 H.R. Aroa, *Transports Canada*
 M. Bankier, *Statistique Canada*
 *K.G. Basavarajappa, *Statistique Canada*
 G. Brackstone, *Statistique Canada*
 D. Bellhouse, *University of Western Ontario*
 L. Biggert, *Université de Florence*
 *D.A. Binder, *Statistique Canada*
 R.D. Burgess, *Statistique Canada*
 C.-M. Cassel, *Statistics Sweden*
 N. Chinnappa, *Statistique Canada*
 *G.H. Choudhry, *Statistique Canada*
 D. Dodds, *Statistique Canada*
 D. Drew, *Statistique Canada*
 M. Eagen, Goss, Gilroy and Associates
 J.L. Eltinge, *Iowa State University*
 I.P. Fellegi, *Statistique Canada*
 W.A. Fuller, *Iowa State University*
 J.F. Gentleman, *Statistique Canada*
 E. Gbur, *University of Arkansas*
 *G.B. Gray, *Statistique Canada*
 M.A. Hidiroglou, *Statistique Canada*
 D. Holt, *University of Southampton*
 S. Ingram, *Statistique Canada*
 *K.M. Wolter, *U.S. Bureau of the Census*
 W.D. Kalsbeek, *University of North Carolina*
 *G. Kalton, *University of Michigan*
 N.J. Kirkendall, *Alexandria, Virginia*
 G. Kriger, *Statistique Canada*
 H. Lee, *Statistique Canada*
 J.M. Lepkowski, *University of Michigan*
 *S. Kumar, *Statistique Canada*
 R. Lachapelle, *Statistique Canada*
 E. Langel, *Statistique Canada*
 S. Michaud, *Statistique Canada*
 D.G. Paton, *Statistique Canada*
 M. Podehl, *Statistique Canada*
 *J.N.K. Rao, *Carleton University*
 P.S.R.S. Rao, *University of Rochester*
 G. Sande, *Statistique Canada*
 *I. Sande, *Statistique Canada*
 C.E. Särndal, *Université de Montréal*
 *F. Scheuren, *U.S. Internal Revenue Service*
 E.A. Schillmoeller, *Nielsen Media Research*
 M. Sheridan, *Statistique Canada*
 A.C. Singh, *Statistique Canada*
 K.P. Srinath, *Statistique Canada*
 V. Tremblay, *Statplus*
 A. van Baaren, *Statistique Canada*

On remercie également ceux qui ont contribué à la production des numéros de la revue pour 1987: B. Babcock (Rédaction), C. VanBastelaar (Photocomposition), G. Gaulin (Services aux auteurs) et M. Haight (Services de traduction).
 On tient à remercier le personnel des Divisions des méthodes d'enquêtes sociales et des méthodes d'enquêtes-entreprises qui ont aidé à la correction et à la vérification. Finalement on désire exprimer notre reconnaissance à J. Clarke, E. Corriveau, J. Dufrêne, M. Kent, C. Larabie, D. Lemire et N. Smalldridge pour leur apport à la coordination, la dactylographie et à la rédaction.

- RAO, J.N.K. (1965). On two simple schemes of unequal probability sampling without replacement. *Journal of the Indian Statistical Association*, 3, 173-180.
- SANKARANARAYANAN, K. (1969). An IPPS sampling scheme using Lahiri's method of selection. *Journal of the Indian Society of Agricultural Statistics*, 21, 58-66.
- SEN, A.R. (1952). Further developments of the theory and application of primary sampling units with special reference to the North Carolina agricultural population. Thèse de doctorat, North Carolina State College, Raleigh.

et dans ce cas la méthode de Horvitz-Thompson est supérieure à la méthode de Hansen-Hurwitz. La condition définie en (4) a été établie pour la première fois par Rao (1963) lors-que $n = 2$ et que la méthode de Midzuno-Sen est utilisée, mais il est intéressant de constater par l'équation (5) que la condition ne change pas même pour des valeurs de n supérieures à 2. Chaudhuri (1975) et Mukhopadhyay (1975) ont, chacun de leur côté, obtenu les résultats ci-dessus pour la méthode de Midzuno-Sen.

Brewer (1963), Rao (1965) et Durbin (1967) ont proposé diverses méthodes d'échantillon-nage avec PPT pour $n = 2$ avec les mêmes probabilités d'inclusion,

$$\pi_{ij} = \frac{1+k}{2p_i p_j} \left(\frac{1}{1-2p_j} + \frac{1}{1-2p_i} \right) \text{ où } k = \sum_{i=1}^N \frac{1-2p_i}{p_i}.$$

Ces méthodes ne sont pas assujéties aux conditions qui s'appliquent aux méthodes précédentes. Par (3),

$$\phi = \frac{1+k}{N} \sum_{i=1}^N \frac{1-2p_i}{Z_i^2} \geq 0,$$

de sorte que la méthode de Horvitz-Thompson est de nouveau supérieure à la méthode de Hansen-Hurwitz.

REMERCIEMENTS

L'auteur tient à exprimer sa reconnaissance au rédacteur en chef, M. P. Singh, et à l'ar-bitre pour leurs nombreux commentaires utiles.

BIBLIOGRAPHIE

- BREWER, K. R. W. (1963). A model of systematic sampling with unequal probabilities. *Australian Jour-nal of Statistics*, 5, 5-13.
- CHAUDHURI, A. (1975). On some properties of the sampling scheme due to Midzuno. *Bulletin of Calcutta Statistical Association*, 23, 1-19.
- DURBIN, J. (1967). Design of multi-stage surveys for the estimation of sampling errors. *Applied Statistics*, 16, 152-164.
- HANSEN, M. H., et HURWITZ, W. N. (1943). On the theory of sampling from finite population. *An-nals of Mathematical Statistics*, 14, 333-362.
- HORVITZ, D. G., et THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- MIDZUNO, H. (1952). On the sampling system with probability proportionate to sum of sizes. *An-nals of Institute of Statistical Mathematics*, 3, 99-107.
- MUKHOPADHYAY, P. (1975). PPS sampling schemes to base HTE. *Bulletin of Calcutta Statistical Association*, 23, 21-44.
- RAO, J. N. K. (1963). On two systems of unequal probability sampling without replacement. *Annals of Institute of Statistical Mathematics*, 15, 67-72.

où π_i ($i = 1, 2, \dots, N$) désigne la probabilité d'inclusion de la i -ième unité de population U_i dans l'échantillon. En outre, lorsque π_i est proportionnelle à X_i , nous parlons d'une méthode d'échantillonnage avec PIPT. Pour ce genre de méthode,

$$\text{Var}(t_{HT}) = \sum_{i=1}^N \frac{U_i^2}{Z_i^2} + \sum_{i \neq j=1}^N Z_i Z_j \frac{U_i U_j}{\pi_{ij}} \quad (2)$$

où Z_i est défini en (1) et π_{ij} ($i \neq j = 1, 2, \dots, N$) est la probabilité (conjointe) d'inclusion des i -ième et j -ième unités de population dans l'échantillon. Lorsqu'on définit une méthode d'échantillonnage avec PIPT, π_{ij} peut être exprimée de façon plus simple.

Par (1) et (2),

$$\phi = \text{Var}(t_{HT}) - \text{Var}(t_{HH}) = \sum_{i \neq j=1}^N Z_i Z_j \frac{U_i^2 U_j}{\pi_{ij}} \quad (3)$$

2. COMPARAISON DES MÉTHODES

Pour estimer T_y , Midzuno (1952), Sen (1952) et Sankaranarayanan (1969) ont proposé des méthodes d'échantillonnage avec PIPT fondées sur l'estimateur d'Horvitz-Thompson t_{HT} . La méthode de Midzuno et Sen est applicable si

$$p_i = \frac{X_i}{T_x} > \frac{n(N-1)}{n-1}, i=1, \dots, N, \quad (4)$$

tandis que celle de Sankaranarayanan est applicable si

$$\sum_{j \in S} p_j > (n-1)/(N-1) \text{ pour tous } s \in S,$$

qui est une condition moins stricte. Pour les deux méthodes, les probabilités conjointes d'inclusion sont définies par

$$\pi_{ij} = \frac{n(n-1)}{n(n-1)} (p_i + p_j - \frac{N-1}{1}).$$

Ainsi, par l'équation (3),

$$\phi = \frac{n(n-1)}{n(n-1)} \left[\sum_{i=1}^N \frac{U_i^2}{Z_i^2} \left(2 - \frac{(N-1)p_i}{1} \right) + \frac{(N-1)}{1} \left(\sum_{i=1}^N \frac{U_i}{Z_i} \right)^2 \right] \quad (5)$$

L'expression ci-dessus est non-négative si

$$p_i > \frac{2(N-1)}{1}, i=1, 2, \dots, N,$$

Comparaison de la méthode de Horvitz-Thompson et de la méthode de Hansen-Hurwitz

S.G. PRABHU-AJGAONKAR¹

RÉSUMÉ

La méthode de Hansen-Hurwitz (1943) est réputée moins efficace que la méthode de Horvitz-Thompson (1952) qui est liée à un certain nombre de méthodes d'échantillonnage avec PPT (probabilité d'inclusion proportionnelle à la taille). Le présent article démontre de façon simple la supériorité de la seconde méthode et présente donc un intérêt au point de vue pédagogique.

MOTS CLÉS: Méthodes d'échantillonnage; probabilité d'inclusion proportionnelle à la taille; forme quadratique définie positive.

1. INTRODUCTION

Soit U une population finie constituée de N unités identifiables $[U_1, U_2, \dots, U_N]$. Deux nombres X_i et Y_i se rattachent à U_i , la i -ième unité de la population; les valeurs de X_i sont connues tandis que celles de Y_i sont inconnues mais fixes. De façon générale, X_i est une mesure de la taille de U_i qui est fortement corrélée avec Y_i .

Si nous utilisons la méthode de Hansen-Hurwitz (1943) pour estimer le total de la population $T_y = Y_1 + Y_2 + \dots + Y_N$, nous devons choisir avec remise n unités de population avec probabilité proportionnelle à X_i et utiliser l'estimateur sans biais

$$t_{HH} = \frac{1}{n} \sum_{r=1}^n p_r Y_r,$$

où $p_r = X_r / T_x$, $T_x = X_1 + X_2 + \dots + X_N$, et Y_r ($r=1, 2, \dots, n$) est le résultat du r -ième prélèvement. Étant donné que $\sum Z_i = 0$, il est facile de montrer que

$$\text{Var}(t_{HH}) = \sum_{i=1}^I \frac{Z_i^2}{n p_i}, \quad (1)$$

où $Z_i = Y_i - p_i T_y$, $i=1, 2, \dots, N$.

Dans le cas d'un échantillonnage sans remise, Horvitz et Thompson (1952) ont proposé l'estimateur sans biais

$$t_{HT} = \frac{\sum_{i=1}^I \pi_i}{n},$$

¹ S.G. Prabhu-Ajgaonkar, Department of Mathematics and Statistics, Marathwada University, Aurangabad 431004, Inde.

- Oh et Scheuren: Variante de la méthode itérative du quotient
- KULLBACK, S. (1968). *Information Theory and Statistics*. New York: Dover.
- LESZCZ, M., OH, H.L., et SCHEUREN, F. (1983). Modified raking estimation in the corporate SOI program. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 107-111.
- OH, H.L., et SCHEUREN, F.J. (1978a). Multivariate raking ratio estimation in the 1973 exact match study. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 716-722.
- OH, H.L., et SCHEUREN, F.J. (1978b). Some unresolved application issues in raking ratio estimation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 723-728.
- OH, H.L., et SCHEUREN, F.J. (1983). Weighting adjustments for unit nonresponse. Dans *Incomplete Data in Sample Surveys, Volume 2, Theory and Bibliographies*, (éd. W.G. Madow, I. Olkin et D.B. Rubin), New York: Academic Press, 143-184.
- RAO, J.N.K. (1985). Inférence conditionnelle dans les enquêtes par sondage. *Techniques d'enquête*, volume 11, 15-31.
- SCHEFFÉ, H. (1959). *The Analysis of Variance*. 106-119. New York: John Wiley.
- SCHEUREN, F., SCHWARTZ, O., et KILSS, B. (1984). Statistics from individual income tax returns: quality issues and budget cut impact. *Review of Public Data Use*, 12, 55-67.
- STEPHAN, F.F. (1942). Iterative method of adjustment sample frequency tables when expected margins are known. *Annals of Mathematical Statistics*, 13, 166-178.
- U. S. DEPARTMENT OF TREASURY (1984). Internal Revenue Service, *Statistics of Income - 1980 Corporation Income Tax Returns*, publication n° 16, U.S. Department of Treasury.
- U. S. DEPARTMENT OF TREASURY (1987). Internal Revenue Service, *Statistics of Income - 1984 Corporation Income Tax Returns*, publication n° 16, U.S. Department of Treasury.
- WESTAT, INC. (1974). Results of a study to improve sampling efficiency of statistics of corporation income. Bethesda, Maryland, (non publié).

sociétés, qui a été habilement secondé par Richard Collins. Nat Shaifer a fourni des données éclairantes sur le degré de comparabilité de la classification industrielle utilisée à des fins administratives et de celle utilisée à des fins statistiques. Mike Leszcz a joué un rôle de premier plan en exposant les résultats préliminaires de notre étude à l'assemblée de l'American Statistical Association de 1983 tenue à Toronto. Les auteurs tiennent aussi à exprimer leur reconnaissance à Rod Little, qui par des échanges a contribué à améliorer la qualité de cet article, et aux arbitres, qui ont apporté de nombreuses suggestions utiles. Enfin, Bettye Jamerson a secondé les auteurs dans la préparation du manuscrit.

BIBLIOGRAPHIE

- BANKIER, M.D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association*, 81, 1074-1079.
- BINDER, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.
- BRACKSTONE, G.J., et RAO, J.N.K. (1979). An investigation of raking ratio estimators. *Sankhya*, Sér. C, 41, 97-114.
- CAUSEY, B.D. (1972). Sensitivity of raked contingency table totals to changes in problem conditions. *Annals of Mathematical Statistics*, 43, 656-658.
- COCHRAN, W.G. (1977). *Sampling Techniques*, (3^e éd.). New York: John Wiley.
- DEMING, W.E. (1943). *Statistical Adjustment of Data*. New York: Dover.
- DEMING, W.E., et STEPHAN, F.F. (1940). On a least square adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11, 427-444.
- FAN, M.C., WOLTMAN, H.F., MISKURA, S.M., et THOMPSON, J.H. (1981). 1980 Census variance estimation procedure. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 176-181.
- FIENBERG, S.E. (1970). An iterative procedure for estimation in contingency tables. *Annals of Mathematical Statistics*, 907-917.
- FULLER, W.A. (1966). Estimation employing post strata. *Journal of the American Statistical Association*, 61, 1172-1183.
- HANSON, R. (1978). The Current Population Survey: design and methodology. Technical Paper No. 40, U. S. Bureau of the Census.
- HINKINS, S., JONES, H., et SCHEUREN, F. (1987). Mise à jour des probabilités de sélection des déclarations d'impôt dans le cadre du programme de la statistique du revenu des sociétés. Communication présentée au Symposium international sur les utilisations statistiques des données administratives, Ottawa.
- HINKINS, S., et SCHEUREN, F. (1986). L'imputation par la méthode "hot deck" appliquée à un plan d'échantillonnage à deux degrés. *Techniques d'enquête*, 12, 181-195.
- HOLT, D., et SMITH, T. M. F. (1979). Post-stratification. *Journal of the Royal Statistical Society, Sér. A*, 142, 33-46.
- IRELAND, C.T., et KULLBACK, S. (1968). Contingency tables with given marginals. *Biometrika*, 55, 179-188.
- IRELAND, C.T., et SCHEUREN, F.J. (1975). The rake's progress. Dans *Computer Programs for Contingency Table Analysis*, D.C.: The George Washington University, 155-216.
- JONES, H., et MCMAHON, P.B. (1984). Sampling corporation income tax returns for Statistics of Income, 1951 to present. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 269-274.

nous avons groupé les données des cellules à fréquence moins élevée, y compris le calcul de la moyenne des facteurs de pondération W_{ij} qui en résulte et l'estimation des V_{ij} fondée sur le regroupement de cases. Le fait d'avoir totalisé les données selon la classification industrielle destinée à des fins statistiques au lieu de celle destinée à des fins administratives, comme c'est le cas dans le tableau I, peut aussi expliquer largement ces différences.

4. CONCLUSIONS ET SUJETS DE RECHERCHE

4.1 Généralités

La version modifiée de la méthode itérative du quotient pour l'échantillon de sociétés représente sûrement un progrès par rapport à l'ancienne méthode de l'échantillonnage stratifié normal. Cependant, il y a encore un certain nombre de points qui nous agacent. Par exemple, il y aurait lieu d'approfondir le lien qui existe entre les méthodes classiques de regroupement et la version modifiée de la méthode itérative du quotient. Toutefois, il ne suffit pas pour cela d'examiner les modifications que l'on peut apporter aux méthodes d'estimation. Il semble tout aussi important de se pencher sur le plan de sondage fondamental. Enfin, l'approximation de la variance utilisée dans le présent article mérite d'être approfondie. Le fait d'avoir opté pour la stabilité et la simplicité des calculs peut nous avoir coûté cher. Comme nous l'avons dit plus tôt, les ouvrages de statistique regorgent de bonnes solutions qui méritent d'être comparées sous tous leurs aspects avec ce qui se fait présentement.

4.2 Questions relatives à l'estimation

Intuitivement, nous sommes fortement tentés de mettre au point une méthode de stratification a posteriori qui nous permettrait de modifier *graduellement* les relations conditionnelles basées sur les fréquences marginales pour qu'elles soient basées également sur les fréquences de cellules. L'inconvénient de la méthode que nous utilisons actuellement est qu'elle a un aspect *temporaire*. De fait, cette méthode n'est qu'un pis-aller en attendant que nous puissions accroître la qualité des données administratives des industries. Notre principal souci est de réduire la variation de réponse qui découle des erreurs de déclaration ou de traitement. Même si nous ne réussissons pas à accroître directement la qualité des données administratives, nous pouvons atténuer les effets de l'erreur de réponse en examinant les tableaux selon l'industrie et la strate d'échantillon sur plusieurs années. C'est ce que nous projetons de faire et c'est ce qui nous permettra peut-être de mieux intégrer d'une part la méthode itérative du quotient et d'autre part l'estimation par la stratification a posteriori fondée sur le regroupement de cellules.

4.3 Questions relatives au plan de sondage

Des données administratives de meilleure qualité pour les industries ont une utilité évidente dans la conception du plan de sondage. À l'heure actuelle, les coefficients de variation diffèrent très largement d'une industrie à l'autre et les industries de moindre importance sont très faiblement représentées. La stratification a posteriori, aussi poussée soit-elle, ne peut corriger à elle seule la situation. Une meilleure répartition de l'échantillon entre les industries, et dans le temps, semble compter parmi les mesures les plus urgentes (voir Hinkins, Jones et Scheuren 1987).

REMERCIEMENTS

Comme c'est le cas pour presque tous les travaux de recherche appliquée, les auteurs tiennent à remercier les nombreuses personnes qui ont rendu possible cet article. Ils veulent tout d'abord souligner la contribution notable d'Homér Jones, responsable de l'échantillon des

Réduction de l'erreur quadratique moyenne (EQM) en pourcentage de l'EQM correspondante pour l'échantillonnage stratifié normal

Estimateur	Nombre de déclarations			Recettes totales			Crédit d'impôt à l'emploi		
Estimateurs de la méthode itérative du quotient "pure":	PRRE	98.6	8.3	98.6	9.2	- 3.09	PRRE	- 3.09	+ 1.09
	PRRE (400)	98.6		98.6	9.2	- 3.09	PRRE (400)	- 3.09	+ 1.09
	PRRE (200)	98.6		98.6	11.9	- 3.09	PRRE (200)	- 3.09	+ 1.09
	Estimateurs de la méthode itérative du quotient bornée:								
Estimateurs de la méthode itérative du quotient bornée:	BRRE	74.0	13.8	74.0	15.6	+ 1.09	BRRE	+ 1.09	+ 1.09
	BRRE (400)	73.4		73.4	15.6	+ 1.09	BRRE (400)	+ 1.09	+ 1.09
	BRRE (200)	72.3		72.3	17.4	+ 1.09	BRRE (200)	+ 1.09	+ 1.09

Remarque: Les pourcentages indiqués sont des moyennes simples des réductions (en pourcentage) observées dans les 56 grandes branches d'activité ayant servi à la stratification a posteriori. On remarquera que les pourcentages indiqués dans la partie supérieure de la première colonne (nombre de déclarations) ne sont pas tout à fait égaux à 100 parce que l'estimation itérative par le quotient s'est appliquée à toutes les sociétés et que les N_{ij} et n_{ij} ont été définies en conséquence tandis que le tableau porte uniquement sur les sociétés actives (environ 90%). Les pourcentages indiqués dans la partie inférieure de la première colonne (nombre de déclarations) diffèrent l'un de l'autre et des pourcentages indiqués dans la partie supérieure de cette colonne parce que l'estimation itérative ne s'est pas faite jusqu'à la convergence. Cela n'est plus le cas depuis l'année d'imposition 1985.

nombre de ces industries à 56 par la technique du regroupement. Malgré cela, nous avons encore un tableau de taille très appréciable (1,568 cellules).
Il convient de souligner que l'univers comptait 414 cellules à fréquence nulle à l'origine et que l'échantillon en a généré 125 autres. Avant l'estimation itérative par le quotient, nous avons extrait 96 cellules qui renfermaient 200 observations de l'échantillon ou plus et les avons corrigées séparément par la méthode du quotient. (En tout, 57% de l'échantillon de formules 1120 et 1120S a été corrigé de cette façon). Enfin, il a fallu borner 73 autres cellules au cours de l'estimation itérative par le quotient. Cela signifie que 708 cellules en tout (45%) ont été considérées comme des cellules à fréquence nulle durant l'estimation itérative par le quotient. On a amorcé l'estimation itérative par le quotient en appliquant l'estimateur stratifié normal à chaque cellule du tableau. On a utilisé deux séries de comparaisons faites ici, l'estimation-industrie et période d'échantillonnage et 2) par strate d'échantillon et période d'échantillonnage. Dans les données publiées pour 1984 et dans les comparaisons faites ici, l'estimation itérative par le quotient n'a pas été effectuée jusqu'à la convergence; le processus s'est limité en effet à deux cycles. (À ce propos, l'incertitude qui entourait le biais conditionnel de cette méthode nous a amené à effectuer l'estimation itérative par le quotient jusqu'à la convergence pour les données de l'échantillon de 1985.)

Les travaux de 1984 ont eu pour effet de réduire sensiblement l'erreur type des fréquences pour l'ensemble des industries et pour chacune d'elles, les réductions étant toutefois environ à moitié moindres que celles indiquées dans le tableau 1. On a aussi observé des réductions modérées dans le cas des recettes totales (8,7%) tandis que pour de nombreuses variables comme le crédit d'impôt à l'emploi et le bénéfice net, l'erreur type globale est demeurée à peu près la même (voir U.S. Department of Treasury 1987, pour plus de renseignements). Comme nous l'avons déjà souligné, ces différences de résultats peuvent être attribuées en partie à la nature des relations conditionnelles utilisées (Holt et Smith 1979). Les résultats initiaux dépendaient des relations conditionnelles basées sur les fréquences marginales de l'échantillon n_i et n_j tandis que les résultats ultérieurs reposaient sur des relations conditionnelles plus complexes. Nous cherchons encore d'autres raisons qui pourraient expliquer pourquoi les réductions d'erreur type observées en 1984 sont moins fortes que celles observées dans les études antérieures. Une des raisons qui nous viennent immédiatement à l'esprit est la manière dont

en cinq étapes: 1) formation des demi-échantillons; 2) classification des données des demi-échantillons selon deux critères – selon la strate d'échantillon originale et selon l'industrie principale (strate formée a posteriori); 3) calcul d'une série de poids pour chaque demi-échantillon et pour chaque estimation; 4) calcul de l'estimation de certaines unités par l'application de poids à des données d'échantillon produites par chaque demi-échantillon. Pour l'EQM fondé sur les variations des estimations produites par chaque demi-échantillon. Pour des raisons monétaires, nous n'avons utilisé que 14 ensembles de demi-échantillons. Les résultats de cette comparaison sont présentés sommairement dans le tableau 1; les chiffres concernant le nombre de déclarations sont ce à quoi on s'attendait. Des réductions de près de 100% ont été observées pour les estimations PRRF, PRRF(200), et PRRF(400). Le fait d'appliquer les bornes $\sqrt{2/3}$ et $\sqrt{3/2}$ et de se limiter à deux cycles s'est traduit par des réductions moins fortes mais tout de même appréciables. Le tableau 1 indique aussi des réductions de l'EQM pour les recettes totales, une variable clé, mais ces réductions n'ont pas l'ampleur des premières.

Nous avons inclus dans le tableau 1 les résultats relatifs au crédit d'impôt à l'emploi pour illustrer l'ambivalence qui peut exister dans le cas des variables qui ne sont pas liées étroitement à l'industrie. En particulier, nous observons des hausses (modestes) de l'erreur quadratique moyenne pour cette variable du fait, probablement, que le crédit d'impôt à l'emploi est moins dépendant de la classification des industries utilisée dans cette étude. En ce qui concerne les recettes totales, il convient de souligner que les réductions de l'erreur quadratique moyenne (attribuables au remplacement de l'estimation initiale (NORM) par l'estimation itérative par le quotient se comparent toutes avantageusement aux résultats de l'étude pilote de Westat. Même si cette constatation est encourageante, la situation a beaucoup évolué entre le moment où Westat a fait connaître les résultats de son étude et la parution de l'ouvrage de Leszcz, Oh et Schuereen (1983). Si nous voulions faire une comparaison vraiment révélatrice, à laquelle nous n'avons pas songé jusqu'à maintenant, il suffirait d'appliquer les méthodes classiques de regroupement et notre version modifiée de la méthode itérative du quotient à la même série de données.

Nous nous permettons ici une dernière remarque à propos du tableau 1; les chiffres de ce tableau ont été établis à partir des données administratives des industries qui ont servi à la stratification a posteriori. Comme la classification des industries n'est pas la même pour les besoins administratifs et les besoins statistiques publiés puisque le code d'une forte proportion des entités (plus de 20%) est modifié au cours du traitement de l'échantillon des sociétés.

3.3 Version modifiée de la méthode itérative du quotient – résultats courants

Depuis l'année d'imposition 1980, nous produisons et publions régulièrement les statistiques des sociétés à l'aide de la méthode itérative du quotient bornée BRRE (200) (U.S. Department of Treasury 1984). Pour les années d'imposition 1983 et suivantes, nous avons fait les modifications décrites dans la sous-section 2.3 de manière à pouvoir calculer des variances conditionnelles approximatives. L'année d'imposition 1984 a été la première pour laquelle de telles variances ont été publiées (U.S. Department of Treasury 1987). De plus, afin de vérifier les résultats antérieurs, nous avons comparé pour l'année d'imposition 1984 la variance conditionnelle de la version modifiée de la méthode itérative du quotient avec la variance qui aurait été estimée si nous avions eu recours à l'échantillonnage stratifié normal. Avant d'analyser les résultats de cette comparaison, il serait utile de fournir quelques précisions sur le modèle utilisé pour 1984.

Dans un ouvrage antérieur (Leszcz, Oh et Schuereen 1983) et pour 1984, l'univers des formules 1120 et 1120S des déclarations des sociétés a été réparti entre 58 grandes branches d'activité. Pour 1984, chaque branche d'activité a été subdivisée en 14 strates d'échantillon à chacune des deux années durant lesquelles l'échantillon devait être choisi. Quelques-unes des principales branches d'activité étaient si clairsemées que nous avons réduit sans tarder le

fréquence, Vestat a utilisé les méthodes classiques de regroupement pour combiner des strates formées a posteriori selon l'industrie à l'intérieur des strates d'échantillon existantes. On a continué à se préoccuper de la qualité des données administratives au niveau de l'industrie, particulièrement en ce qui a trait aux cas à faible fréquence; quoiqu'il en soit, la méthode Vestat n'a jamais été mise en application à cause de questions opérationnelles plus pressantes. Une importante série de compressions budgétaires imposées durant la période 1980-1982 ont entraîné une modification des plans de sondage et des méthodes d'estimation pour pres-que tous les projets qui constituent le programme de la statistique du revenu (voir, par exem-ple, Hinkins et Scheuren 1986; Scheuren, Schwartz et Kilss 1984); on a notamment réduit la taille des échantillons pour l'enquête sur les sociétés et cette mesure, quoique partielle-ment annulée ultérieurement, a contribué à rouvrir le débat sur la stratification a posteriori par industrie.

L'application de la méthode itérative du quotient semblait présenter plus d'intérêt que les travaux de Westat parce qu'on n'était pas tant préoccupé par la qualité des fréquences marginales pour l'industrie que par celle des fréquences de cellules. En outre, il n'était plus nécessaire de mettre en application une méthode de regroupement.

3.2 Version modifiée de la méthode itérative du quotient - résultats préliminaires

Lorsque nous avons appliqué la version pure de la méthode itérative du quotient à l'échan-tillon de l'année d'imposition 1979, nos principaux clients nous ont fait part de leurs in-quiétudes à ce sujet. Ils craignaient surtout que des facteurs de correction élevés puissent avoir un effet négatif sur certaines statistiques. Nous-mêmes, à la lecture des résultats, craig-nions de ne pas avoir réalisé nos objectifs en ce qui concerne les combinaisons "industrie-strate d'échantillon" pour lesquelles le nombre d'observations de l'échantillon était élevé. En conséquence, ces résultats n'ont jamais été utilisés et les données de l'année d'imposition 1979 ont été publiées sur la base d'un échantillonnage stratifié normal (NORM).

Les recherches se sont poursuivies toutefois et en 1983, on en est arrivé à comparer entre elles les erreurs quadratiques moyennes de six variantes de la méthode itérative du quotient et à les comparer aux résultats de nos études antérieures (Leszcz, Oh et Scheuren 1983). Ces variantes comportaient trois méthodes "pures":

PRRE: Méthode itérative du quotient "classique" avec convergence (Deming et Stephan 1940).

PRRE (200): Correction par le quotient pour les cellules avec échantillons de 200 déclara-tions ou plus et méthode itérative du quotient "classique" pour les cellules restantes avec convergence.

PRRE (400): Correction par le quotient pour les cellules avec échantillons de 400 déclara-tions ou plus et méthode itérative du quotient "classique" pour les cellules restantes avec convergence.

En outre, il y avait trois versions de l'estimation itérative par le quotient bornée, où les bornes étaient $(\sqrt{2/3}, \sqrt{3/2})$. Ces trois versions étaient:

BRRE: Estimation itérative par le quotient bornée (2 cycles).
BRRE (200): Correction par le quotient pour les cellules avec échantillons de taille $n = 200$ et estimation itérative par le quotient bornée (2 cycles) pour les cellules restantes.

BRRE (400): Correction par le quotient pour les cellules avec échantillons de taille $n = 400$ et estimation itérative par le quotient bornée (2 cycles) pour les cellules restantes.

En ce qui concerne l'estimation itérative par le quotient bornée, nous n'étions pas sûrs au départ que la convergence complète était possible; nous avons donc effectué une simplifica-tion et n'avons soumis les équations conditionnelles (2.2) et (2.3) qu'à deux cycles. Pour comparer les erreurs quadratiques moyennes (EQM), nous avons tiré des pseudo-échantillons répétés conçus de la même façon que l'échantillon général. Cela s'est fait

Elle est comparable en ce sens que nous bornons le \tilde{W}_{ij} se rapportant à la cellule qui doit être restreinte, puis extrayons respectivement des tableaux de la population et de l'échantillon la fréquence estimée $\tilde{N}_{ij} = W_{ij} n_{ij}$ et la fréquence n_{ij} pour la cellule en question, pour enfin corriger les observations restantes.

Ces "solutions" partielles soulèvent trois problèmes particuliers. Il y a tout d'abord la définition plutôt arbitraire d'une cellule à fréquence "élevée" et d'un coefficient de pondération qui s'écarte "trop" de sa valeur initiale. En deuxième lieu, si nous étions disposés à utiliser l'estimation ordinaire par le quotient pour les cellules à fréquence "élevée", il faudrait savoir pourquoi ne pouvons-nous pas utiliser les méthodes classiques de regroupement pour les cellules restantes. Le troisième problème a trait aux propriétés de la convergence de l'algorithme itératif lorsque nous utilisons la méthode hybride. Par exemple, les études qui ont été faites sur l'estimation itérative par le quotient montrent très clairement que les tableaux qui renferment un trop grand nombre de zéros seront très instables et que le processus d'itération pourrait ne pas converger (voir, par exemple, Oh et Scheuren 1978b; Ireland et Scheuren 1975). Ce problème revêt un intérêt particulier puisque les modifications proposées ont justifié ment pour but d'introduire des zéros dans le tableau. Si ces zéros sont *mal distribués*, cela pourrait avoir des conséquences très néfastes pour le taux de convergence et même pour la qualité des estimateurs. Nous nous sommes donc proposé au départ d'appliquer la méthode hybride relativement peu souvent. Le présent article n'a pas pour but de résoudre les problèmes évoqués ci-dessus (si ceux-ci peuvent être effectivement résolus.) Néanmoins, nous y reviendrons à la section 3, où nous les réexaminerons en fonction du modèle d'application utilisé, et à la section 4, où nous proposerons des sujets de recherche.

3. ESTIMATION ITÉRATIVE PAR LE QUOTIENT DANS LE CADRE DU PROGRAMME DE LA STATISTIQUE DU REVENU DES SOCIÉTÉS

3.1. Contexte

Depuis plus de 70 ans, l'Internal Revenue Service des États-Unis produit chaque année des statistiques fondées sur les déclarations d'impôt des sociétés. De fait, les données des sociétés constituent l'élément essentiel du programme de la statistique du revenu, qui est l'application donnée à l'ensemble des séries de données non administratives produites par l'Internal Revenue Service à l'intention du public.

Jusqu'en 1951, les statistiques des sociétés étaient fondées sur un dénombrement complet des déclarations produites. Depuis lors, on utilise un échantillon aléatoire stratifié formé d'environ 90,000 déclarations annuellement (environ 3,000,000 déclarations sont faites chaque année). L'actif et le revenu sont les principales variables de stratification (Jones et McMahon 1984). La stratification par industrie a été longtemps envisagée mais la qualité des codes d'industrie inscrits par les contribuables ne semblait pas assez bonne pour justifier un usage généralisé de ce mode de stratification. Au niveau d'aggrégation le plus détaillé, par exemple, 20% et plus des codes inscrits par les contribuables sont modifiés durant le traitement statistique. Néanmoins, vu l'importance des statistiques des industries pour stratifier l'échantillon à de tenter d'utiliser les données administratives des industries pour stratifier l'échantillon à posteriori et c'est ce qu'on a essayé de faire pendant de nombreuses années (voir, par exemple, Westat Inc. 1974; Leszcz, Oh et Scheuren 1983).

Dans une étude pilote sur la stratification réalisée par Westat au début des années 1970, on a réussi à réduire de façon notable les erreurs types pour un certain nombre de variables, notamment les recettes totales (pour lesquelles on a observé une réduction de l'erreur type d'environ 12%). Toutefois, des hausses d'erreur type ont été observées pour des variables qui ne sont pas liées étroitement à l'industrie (par exemple, dividendes versés aux actionnaires) mais il s'agissait de hausses mineures. En ce qui concerne les cellules à faible

d'au moins 25 observations chacun ($n_g \geq 25$) chacun modifiera très peu la valeur de l'équation (2.11). En revanche, cela nous permettra de calculer des estimations par la stratification à posteriori de V_{ij} fondées sur le regroupement. C'est la méthode que nous avons utilisée dans la section 3.

Nous nous permettons ici de faire une dernière remarque. La méthode que nous proposons dans cet article est stable et implique des calculs relativement simples. Cependant, le nombre limité d'applications ne permet pas de conclure à l'utilité de la méthode et bien que nous estimions que cette méthode vaut la peine d'être étudiée, nous n'en préconisons aucunement l'application générale pour l'instant.

2.4 Version modifiée de la méthode itérative du quotient

Comme nous l'avons déjà signalé, les \tilde{N}_{ij} sont des MEAN étant donné des conditions relativement générales. Cela ne veut pas dire pour autant que \tilde{X} aura toutes les propriétés d'un MEAN. En effet, s'il n'existe pas une forte corrélation entre la caractéristique Y et les variables utilisées lors de l'estimation itérative par le quotient, la variance de l'estimateur \tilde{Y} pourrait en subir plus fortement les conséquences que celle d'un estimateur ordinaire par le quotient

$$\tilde{Y} \approx \sum_s \left(\frac{N_j}{n_j} \right) \left(\sum_R \sum_{n_{jk}}^i Y_{ijk} \right). \tag{2.12}$$

Il ne semble pas y avoir de solution générale pour résoudre ce dilemme, bien que nous ayons obtenu des résultats modestes dans notre modèle d'application avec deux méthodes qui débordent peut-être le cadre de notre analyse (voir sous-sections 3.2 et 3.3 pour les résultats).

Habituellement, lorsqu'on discute de la méthode itérative du quotient on suppose que les fréquences marginales N_i et N_j sont connues et que les fréquences de cases N_{ij} ne peuvent être estimées qu'à partir de l'échantillon. Dans notre modèle d'application, nous connaissons les fréquences N_{ij} et avons recours à l'estimation itérative par le quotient pour traiter systématiquement les cellules du tableau dont les n_{ij} sont faibles. Nous pouvions évidemment utiliser les méthodes classiques de regroupement (voir, par exemple, Cochran 1977; Fuller 1966) mais celles-ci ne semblaient pas convenir pour des raisons que nous verrons plus tard.

On peut reconnaître que la méthode itérative du quotient est un moyen satisfaisant de traiter les cellules à faible fréquence dans ce contexte mais que dire des cellules à fréquence plus élevée? Dans ce cas, il serait sûrement préférable d'utiliser l'estimateur ordinaire par le quotient. Si tel était le cas, le biais conditionnel pour ces cellules "à forte fréquence" serait nul mais quel effet cela aurait-il sur les autres cellules? Ce raisonnement nous a amené à adopter une méthode d'estimation hybride selon laquelle on utilise l'estimateur ordinaire par le quotient pour les cellules dont les n_{ij} sont élevées. Ces cellules sont ensuite extraites des tableaux de la population et de l'échantillon et les cellules restantes du tableau de l'échantillon sont soumises à l'estimation itérative par le quotient afin de rajuster leurs fréquences en fonction des fréquences marginales corrigées.

En ce qui a trait aux cellules restantes à fréquence très faible, on a appliqué une seconde méthode visant à réduire les effets négatifs que pouvait avoir l'estimation itérative par le quotient sur certaines variables. Nous avons restreint le processus d'itération de manière que les poids W_{ij} ne s'écartent pas "trop" du poids initial. (À propos, cette forme de restriction est courante dans l'estimation classique par le quotient, voir par exemple Hanson 1978). La version restreinte de la méthode itérative du quotient est comparable à la méthode utilisée lorsque seulement une cellule du tableau de l'échantillon renferme une fréquence "élevée".

En utilisant les formules usuelles de l'analyse de variance (voir Scheffé 1959), nous avons

$$(Y_j - \bar{Y}) = (\bar{Y}_i - \bar{Y}) + (\bar{Y}_j - \bar{Y}) + (\bar{Y}_{ij} - \bar{Y}_i - \bar{Y}_j + \bar{Y}); \quad (2.8)$$

le biais conditionnel, étant donné \bar{n} , peut donc être exprimé comme suit:

$$\text{Biais}(\bar{Y} | \bar{n}) = \sum_R^i (\bar{Y}_i - \bar{Y}) (\bar{N}_i - N_i) + \sum_S^j (\bar{Y}_j - \bar{Y}) (\bar{N}_j - N_j) + \sum_S^j \sum_R^i (\bar{Y}_{ij} - \bar{Y}_i - \bar{Y}_j + \bar{Y}) (\bar{N}_{ij} - N_{ij}). \quad (2.9)$$

Si le processus est convergent, les deux premiers termes de la formule du biais conditionnel deviennent nuls. Pour que le troisième terme de la formule devienne nul, qu'il y ait convergence ou non, il suffit que les Y_{ij} soient telles qu'il n'y ait aucune interaction. Dans les grandes enquêtes portant sur de nombreuses variables, cette hypothèse est irréaliste; dans la pratique toutefois, l'interaction est souvent un élément mineur de la décomposition de X_{ij} ; par conséquent, l'estimateur itératif par le quotient peut, dans beaucoup de cas, ne comporter qu'un faible biais même pour des échantillons de taille moyenne.

2.3 Variance conditionnelle

De nombreuses études ont été faites sur la variance conditionnelle et non conditionnelle de l'estimateur itératif par le quotient (voir, par exemple, Binder 1983; Causey 1972; Bankier 1986; Fan *et coll.* 1981; Brackstone et Rao 1979). Dans un de nos ouvrages antérieurs (voir sous-section 3.2), nous avons utilisé des méthodes de répétées d'échantillons (Leszcz, Oh et Scheuren 1983). Ces méthodes (qui équivalaient à définir les conditions en fonction des fréquences marginales de l'échantillon) se sont avérées coûteuses, peu flexibles et relativement instables, ce qui nous a amené à aborder le problème de l'estimation de la variance conditionnelle d'une manière moins élaborée (bien que les conditions étaient plus strictes). Afin de justifier l'approche que nous utilisons, considérons la variance conditionnelle de \bar{Y} , étant donné \bar{n} . En poussant un peu plus loin la recherche entreprise par Oh et Scheuren (1983), il est possible de montrer que

$$\text{Var}(\bar{Y} | \bar{n}) = \sum_R^i \sum_S^j n_{ij} (\bar{W}_{ij})^2 \left(1 - \frac{N_{ij}}{n_{ij}}\right) V_{ij} \quad (2.10)$$

où les V_{ij} sont les variances de la population du ij -ième sous-groupe; de plus, si $N_{ij} = 0$ ou $V_{ij} = 0$ nous posons $V_{ij} = 0$ (nous appliquons aussi la règle $0/0 = 0$ dans l'équation (2.10)). L'équation (2.10) est valide, peu importe que le processus d'itération soit convergent ou non. Malgré cela, cette expression a fait l'objet de peu d'analyses parce qu'on ne peut l'adapter facilement à l'estimation de la variance conditionnelle. Cet inconvénient est surtout attribuable à notre incapacité de calculer des estimateurs stables de V_{ij} lorsque les n_{ij} sont faibles. Pour surmonter la difficulté, nous nous sommes mis à étudier des méthodes de regroupement fondées sur la valeur du poids d'itération. Nous commençons par définir \bar{W}_{ij} comme une approximation de N_{ij}/n_{ij} , ce qui donne

$$\text{Var}(\bar{Y} | \bar{n}) \approx \sum_R^i \sum_S^j n_{ij} \bar{W}_{ij} (\bar{W}_{ij} - 1) V_{ij}. \quad (2.11)$$

Or, si les valeurs de \bar{W}_{ij} sont classées par ordre croissant et qu'elles s'avèrent se situer dans un court intervalle, le fait de calculer la moyenne de ces poids pour des groupes (ordonnés)

soient satisfaites. Chaque étape de l'algorithme débute par les résultats de l'étape précédente, les valeurs N_{ij} ne cessant de varier; le processus se termine lorsqu'un nombre déterminé d'étapes ont été réalisées ou lorsque les expressions (2.2) et (2.3) sont satisfaites simultanément au degré d'approximation voulu. (Voir Oh et Scheuren (1983) pour plus de détails; voir Ireland et Scheuren (1975) pour l'application de cet algorithme à des tableaux multidimensionnels et l'étude des questions relatives aux ressources informatiques.)

Par une application de la théorie de l'information minimum discriminante (Kullback 1968), il est possible de montrer (voir Ireland et Kullback 1968) qu'étant donné certaines conditions de régularité si on ne connaît que les valeurs N_i et N_j , les N_{ij} calculés par un processus convergent sont asymptotiquement non biaisés, suivent une distribution normale et ont une variance minimum (c'est-à-dire, meilleurs estimateurs asymptotiquement normaux ou MEAN). Ces résultats théoriques justifient en partie l'estimateur itératif d'une caractéristique générale d'enquête Y_{ijk} (par exemple, revenu ou actif) lorsque nous cherchons à estimer le total de population

$$Y = \sum_R \sum_S \sum_{N_{ij}}^k Y_{ijk} \tag{2.4}$$

au moyen, par exemple, de la statistique

$$\bar{Y} = \sum_R \sum_S^i \sum_{N_{ij}}^j \bar{n}_{ij} \left(\sum_{n_{ij}}^k Y_{ijk} \right) \tag{2.5}$$

Lors du traitement des données d'enquête, on attribue habituellement un poids

$$\bar{W}_{ij} = \frac{n_{ij}}{N_{ij}} \tag{2.6}$$

à chaque enregistrement du fichier. Il convient de souligner qu'une des caractéristiques de l'algorithme itératif est que N_{ij} est nécessairement égal à 0 si $n_{ij} = 0$. Pour des raisons de commodité, nous poserons $W_{ij} = 0$ dans de telles circonstances.

Dans le reste de cette section, nous allons nous intéresser surtout aux propriétés conditionnelles des divers estimateurs étudiés. Comme le font valoir Holt et Smith (1979) et Rao (1985), cette approche présente un intérêt considérable. (Soit dit en passant, soulignons que Brackstone et Rao (1979), entre autres, ont analysé le comportement conditionnel de l'estimateur itératif. Cependant, les conditions étaient définies en fonction des fréquences marginales de l'échantillon n_i et n_j .)

2.2 Biais conditionnel

À la suite de Oh et Scheuren (1983), nous allons nous arrêter aux propriétés conditionnelles de \bar{Y} , étant donné $\bar{n} = (n_{11}, n_{12}, \dots, n_{RS})$. En particulier, définissons \bar{Y}_{ij} comme la moyenne de la population pour le ij -ième sous-groupe. L'espérance mathématique conditionnelle de \bar{Y} est donc

$$E(\bar{Y} | \bar{n}) = \sum_R \sum_S^i \bar{N}_{ij} \bar{Y}_{ij} = Y + \sum_R \sum_S^i (\bar{Y}_{ij} - \bar{Y}) (N_{ij} - N_{ij}) \tag{2.7}$$

Par conséquent \bar{Y} est conditionnellement biaisé, l'importance du biais variant selon la structure de la population et selon que le processus d'itération converge ou non. (Évidemment, lorsque le processus est convergent, $E(\bar{N}_{ij}) = N_{ij}$ asymptotiquement et non conditionnellement.)

2. ESTIMATION ITÉRATIVE PAR LE QUOTIENT

2.1 Considérations d'ordre général

L'estimation itérative par le quotient suppose habituellement que deux (ou plus de deux) fréquences marginales – par exemple, $N_{i.}$ et $N_{.j}$ – sont connues mais que les fréquences de cellules N_{ij} ne peuvent être estimées qu'à partir de l'échantillon par \widetilde{N}_{ij} ; nous avons ci-dessous une représentation graphique du problème (Deming 1943):

				$N_{.1}$	$N_{.2}$	\dots	$N_{.j}$	\dots	$N_{.s}$	N
				N_{R1}	N_{R2}	\dots	N_{Rj}	\dots	N_{Rs}	N_R
				\vdots	\vdots	\vdots	N_{ij}	\vdots	\vdots	$N_{i.}$
				N_{21}	N_{22}	\dots	N_{2j}	\dots	N_{2s}	N_2
				N_{11}	N_{12}	\dots	N_{1j}	\dots	N_{1s}	N_1
				1	2	\dots	S			

où $i = 1, \dots, R$ et $j = 1, \dots, S$. Le tableau de l'échantillon correspondant est

				$n_{.1}$	$n_{.2}$	\dots	$n_{.j}$	\dots	$n_{.s}$	n
				n_{R1}	n_{R2}	\dots	n_{Rj}	\dots	n_{Rs}	n_R
				\vdots	\vdots	\vdots	n_{ij}	\vdots	\vdots	$n_{i.}$
				n_{21}	n_{22}	\dots	n_{2j}	\dots	n_{2s}	n_2
				n_{11}	n_{12}	\dots	n_{1j}	\dots	n_{1s}	n_1
				1	2	\dots	S			

Lorsqu'on applique l'algorithme d'itération à un échantillonnage aléatoire simple, on commence par poser

$$\widetilde{N}_{ij} = \frac{n}{N} n_{ij},$$

puis on corrige proportionnellement les valeurs \widetilde{N}_{ij} de telle sorte que les équations

$$\sum_S \widetilde{N}_{ij} = N_{i.}$$

et

$$\sum_R \widetilde{N}_{ij} = N_{.j}$$

Variante de la méthode itérative du quotient

H. LOCK OH et FRITZ SCHEUREN¹

RÉSUMÉ

Les auteurs présentent une méthode d'estimation par le quotient qui est une combinaison de la méthode classique et de la méthode itérative du quotient et qui est utilisée lorsque les fréquences de population N_{ij} d'un tableau de contingence sont connues mais qu'un certain nombre des fréquences observées n_{ij} sont faibles ou nulles. Ils décrivent la façon dont cette méthode a évolué dans le cadre du programme de la statistique du revenu des sociétés (Corporate Statistics of Income Program - CSIP) au cours des dernières années. Enfin, ils envisagent d'autres modifications pour l'avenir et en font l'analyse.

MOTS CLÉS: Méthode itérative du quotient; estimation par le quotient; biais et variance conditionnels.

1. INTRODUCTION

La technique d'estimation itérative par le quotient, ou méthode itérative du quotient, est couramment utilisée dans les enquêtes par sondage. Ses applications varient selon la nature du plan de sondage, la quantité d'information supplémentaire disponible et l'importance des erreurs d'observation (attribuables à la non-réponse ou au sous-dénombrement). Deming et Stephan (1940) ont été les premiers à proposer l'utilisation de la méthode itérative du quotient pour établir une correspondance entre les chiffres de population et des données d'échantillon du recensement de la population des États-Unis de 1940. Ils n'ont pas tardé par la suite à approfondir leurs propres recherches (Deming 1943; Stephan 1942). Depuis lors, plusieurs autres ont redéfini à leur façon cette méthode (Fienberg 1970) peut-être parce que intuitivement, l'algorithme d'itération utilisé paraît intéressant.

Cette méthode a été aussi l'objet de nombreux perfectionnements. Par exemple, Ireland et Kuiliack (1968) ont fait d'importantes recherches sur la convergence de l'algorithme. Comme on pouvait s'y attendre, les spécialistes de Statistique Canada et du U.S. Bureau of the Census se sont penchés longuement sur l'application du processus itératif dans les recensements et les sondages, particulièrement lorsque le processus ne va pas jusqu'à la convergence complète. (voir par exemple Brackstone et Rao 1979; Fan et coll. 1981). Oh et Scheuren (1978a) donnent une bibliographie assez complète concernant la recherche statistique qui s'est faite sur l'itération avant 1978.

Habituellement, lorsqu'on discute de la méthode itérative du quotient, on suppose que deux séries (ou plus) de fréquences marginales, disons $N_{i.}$ et $N_{.j}$, sont connues mais que les fréquences de cellules N_{ij} ne peuvent être estimées qu'à partir de l'échantillon. Lorsque les N_{ij} sont aussi connues, on choisit spontanément l'estimateur par le quotient usuel avec des poids N_{ij}/n_{ij} , à moins que la taille des échantillons correspondants n_{ij} ne soit "trop faible".

Le présent article décrit une méthode hybride qui fait intervenir la méthode classique d'estimation par le quotient et la méthode itérative du quotient dans les cas où les fréquences de cellules N_{ij} sont connues mais que certaines des fréquences observées n_{ij} sont faibles (ou nulles). La section 2 renferme une description de la méthode. Dans la section 3, nous commentons certains résultats empiriques produits lors de l'application de la méthode dans le cadre du programme de la statistique du revenu des sociétés (Corporate Statistics of Income Program - CSIP). Dans la section 4, nous faisons une brève rétrospective de la question et proposons des sujets de recherche.

¹ H. Lock Oh et Fritz Scheuren, Statistics of Income Division, Internal Revenue Service, 1111 Constitution Avenue N.W., Washington, D.C. 20224, États-Unis.

RUBIN, D.B. (1976). Inference on missing data. *Biometrika*, 63, 581-592.

SINGH, M.P., DREW, J.D., et CHOUDHRY G.H. (1984). Remaniement de l'enquête sur la population active au Canada à partir des résultats du recensement de 1981. *Techniques d'enquête*, 10, 139-154.

SCHUEREN, F., OH, H.T., VOGEL, L., et YUSKAVAGE, R. (1981). studies from interagency data linkages, report No. 10: methods of estimation for the 1973 exact match study. U.S. Department of Health and Human Services, Social Security Administration, SSA Publication No. 13-11750.

WRIGHT, R.L. (1983). Finite population sampling with multivariate auxiliary information. *Journal of the American Statistical Association*, 78, 879-884.

ZIESCHANG, K.D. (1986). A generalized least squares weighting system for the Consumer Expenditure Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.

ZIESCHANG, K.D. (1987). Sample weighting methods and estimation of totals in the Consumer Expenditure Survey. Document non publié, U.S. Bureau of Labour Statistics.

rapportés ci-dessus donnent à penser qu'il serait possible d'intégrer les deux genres d'estimation (personnes et familles) sans que cela ne réduise pour autant l'efficacité des estimations relatives aux personnes.

5. PROJETS DE RECHERCHE

On procède actuellement à une étude empirique plus poussée des propriétés de l'estimateur par les moindres carrés, où l'on s'intéresse particulièrement à l'évolution chronologique des estimations et à leur efficacité, pour un plus grand nombre de caractéristiques, par rapport aux estimations produites par la méthode itérative du quotient utilisée actuellement dans l'enquête sur la population active. Les résultats ci-dessus donnent à penser que la composition d'âge-sexe d'un ménage a un "pouvoir explicatif" au moins aussi grand que celui du groupe d'âge-sexe même, du moins en ce qui concerne certaines caractéristiques de personnes. Il sera intéressant de voir si les efficacités relatives seront aussi satisfaisantes pour des caractéristiques qui ont une plus forte corrélation avec l'âge et le sexe. Par ailleurs, même si en pratique les poids négatifs sont rares, il faudra penser à élaborer une méthode qui permettra de traiter ces poids le cas échéant. On pourrait, par exemple, les considérer comme des valeurs aberrantes ou peut être les prévenir en limitant la valeur des modifications de poids (Zieschang 1987). Enfin, il serait utile de définir explicitement le modèle de sous-dénombrement sur lequel repose l'estimateur par les moindres carrés pour que l'on puisse faire une évaluation du modèle proprement dit.

REMERCIEMENTS

L'auteur tient à remercier F. Scheuren pour ses commentaires et suggestions lors de la rédaction de cet article.

BIBLIOGRAPHIE

ALEXANDER, C.H. (1987). Une classe de méthodes utilisant des chiffres de population dans la pondération des ménages. *Techniques d'enquête*, 13, 193-210.

AUGER, E. (1987). Family data from the canadian personal income tax file. Dans *Statistics of Income and Related Administrative Record Research: 1986-1987*, (éd. W. Alvey et B. Kilss), Washington, D.C.: Internal Revenue Service, 177-184.

BETHLEHEM, J.C., et KELLER, W.A. (1987). Linear weighting of sample survey data. *Journal of Official Statistics*, 3, 141-153.

FULLER, W.A. (1975). Regression analysis for sample surveys. *Sankhyā*, C37, 117-132.

GOSSSELIN, J.-F., et THÉROUX, G. (1980). Recensement du Canada de 1976 Qualité des données Série I; Sources d'erreurs - Couverture. N° 99-840 au répertoire, Statistique Canada.

KEYFITZ, N. (1957). Estimates of sampling variance where two units are selected from each stratum, *Journal of the American Statistical Association*, 52, 503-510.

OH, H.L., et SCHEUREN, F. (1978). Multivariate raking ratio estimation in the 1973 exact match study. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 716-722.

PAUL, E.C., et LAWES, M. (1982). Caractéristiques des ménages répondants et non répondants dans l'enquête sur la population active du Canada. *Techniques d'enquête*, 8, 48-85.

PLATEK, R., et SINGH, M.P. Méthodologie de l'enquête sur la population active du Canada, N° 71-526 au répertoire, Statistique Canada.

De façon générale, les trois estimateurs, et plus particulièrement A et B, ne produisent pas des estimations très différentes les unes des autres. Le fait d'inclure les enfants dans la pondération semble se traduire par des estimations légèrement plus élevées pour les personnes occupées et les personnes seules et des estimations légèrement moins élevées pour les familles économiques au niveau national et dans les provinces plus importantes (comparer ces résultats à ceux de Scheuren et coll. 1981). Ces chiffres sont conformes aux prévisions bien qu'il subsiste des écarts notables par rapport aux résultats du recensement, qui indiquent (arrondi au millier près) 6,369,000 familles économiques et 2,583,000 personnes seules au pays. Nous pouvons en conclure que même si l'estimateur par les moindres carrés produit des chiffres qui nous rapprochent plus de la réalité (lorsqu'on tient compte des enfants), il nous faudra de l'information supplémentaire précise et récente pour effacer le biais résiduel. L'efficacité de l'estimateur par les moindres carrés n'est pas tout à fait évidente. Certes, si nous devons faire une prévision en nous fondant sur les résultats ci-dessus, nous serions portés à dire que l'estimateur par les moindres carrés sera aussi efficace que l'estimateur par stratification à posteriori en constatant la similitude des estimations produites par les deux genres d'estimateur. Par ailleurs, on devrait s'attendre à des gains d'efficacité pour les estimations relatives aux familles économiques parce que l'estimateur par les moindres carrés fait intervenir les chiffres de population supplémentaires dans le calcul du poids des ménages. Cependant, on ne peut attribuer un poids unique au ménage sans une redistribution des poids. Comme l'indique le tableau 2, la dispersion est un peu plus forte pour les poids calculés par les moindres carrés que pour ceux calculés par les méthodes ordinaires de stratification à posteriori. Le fait de tenir compte des enfants crée une dispersion encore plus forte. Le degré de dispersion des poids reflète essentiellement la disparité qui existe entre la composition de l'échantillon et celle de la population en général au point de vue de l'âge, du sexe et de la taille des ménages. Comme le fait de vouloir calculer un poids unique pour tout le ménage ajoute une contrainte à la méthode d'estimation, on pourrait s'attendre que les variances en subissent quelque peu les conséquences, surtout si aucune autre information supplémentaire n'est utilisée pour l'estimation. Toutefois, la réalité est quelque peu différente. Les variances de l'estimation par stratification à posteriori ont été estimées au moyen de la méthode de Keyfitz (1957), où les échantillons répétés étaient des UPE (unités primaires d'échantillonnage) ou des UPE regroupées. Les variances de l'estimateur par les moindres carrés ont été estimées à l'aide de la méthode décrite dans Fuller (1975). Pour obtenir une meilleure comparabilité, on a calculé les variances pour plusieurs caractéristiques estimées par stratification à posteriori au moyen de la méthode de Fuller et on les a comparées aux variances obtenues par la méthode de Keyfitz. Les deux séries d'estimations étaient très comparables dans tous les cas (1 à 2 pour cent d'écart).

Le tableau 3 donne l'efficacité estimée de l'estimateur par les moindres carrés par rapport à l'estimateur par stratification à posteriori pour les caractéristiques étudiées dans le tableau 1. Les gains d'efficacité pour les estimations relatives aux familles économiques sont substantiels. Les estimations relatives aux personnes occupées et aux personnes seules semblent aussi être légèrement plus efficaces; toutefois, les réductions de la variance pour ces caractéristiques sont faibles, sauf pour les personnes occupées dans la région de l'Atlantique, surtout quand les enfants figurent dans la pondération. Il est intéressant de constater que la taille moyenne d'un ménage dans la région Atlantique est supérieure à ce qu'elle est ailleurs au Canada, bien que l'influence que cela pourrait avoir sur les estimations de personnes occupées n'est pas tout à fait claire. Pour ce qui a trait aux personnes en chômage, la méthode des moindres carrés ne change essentiellement rien aux variances. Il est permis de croire que ces résultats seront observés de façon générale, c'est-à-dire pour des caractéristiques quelconques. Bien que le critère du poids unique par ménage soit restrictif pour les estimations des caractéristiques de personnes, l'estimateur par les moindres carrés semble résoudre cette difficulté grâce aux variables "explicatives" additionnelles du modèle linéaire, c'est-à-dire les moyennes de toutes les variables auxiliaires des ménages. Les résultats préliminaires

Comme il existe des chiffres de population supplémentaires selon l'âge et le sexe pour chaque province, nous avons établi des estimations pour chacune des provinces. Cependant, dans les tableaux qui suivent, les provinces de moindre importance sont fondues en deux groupes.

Tableau 2

Répartition des écarts en pourcentage entre les poids finals et les poids initiaux, estimateur par stratification a posteriori et estimateur par les moindres carrés, enquête sur la population active, mai 1981

Ecart en pourcentage	Pourcentage de l'échantillon		
	Stratification a posteriori	Moindres carrés	Moindres carrés (avec enfants)

> - 30%	0.0	0.1	0.2
- 30 à - 20%	0.0	0.5	0.9
- 20 à - 10%	0.6	3.0	5.3
- 10 à 0%	23.9	20.4	27.1
0 à 10%	53.9	44.6	37.3
10 à 20%	20.6	26.3	21.6
20 à 30%	0.6	4.4	6.2
30 à 40%	0.1	0.4	0.9
40 à 50%	0.0	0.0	0.2
< 50%	0.0	0.0	0.2

Note: La taille de l'échantillon est N = 159014.

Tableau 3

Efficacité estimée de l'estimateur par les moindres carrés par rapport à l'estimateur par stratification a posteriori: nombre de personnes occupées et en chômage, nombre de familles économiques et de personnes seules, enquête sur la population active, mai 1981

	Estimateur ^a	Personnes occupées			Personnes en chômage			Familles économiques			Personnes seules		
		B	C		B	C		B	C		B	C	
Canada		1.044	1.066	0.999	0.999	1.044		1.565	1.616	1.036	1.038	1.048	1.045
Région de l'Atlantique		1.110	1.193	0.977	0.992	1.110		1.266	1.567	0.998	1.070	1.098	1.134
Québec		1.059	1.063	1.005	0.992	1.059		1.553	1.582	1.020	1.064	1.092	1.134
Ontario		1.028	1.059	1.011	1.010	1.028		1.825	1.828	1.064	1.099	1.134	1.134
Région des Prairies		1.001	1.072	1.009	1.066	1.001		1.205	1.420	1.009	1.048	1.070	1.134
Colombie-Britannique		1.038	1.053	0.964	0.978	1.038		1.248	1.203	1.048	1.070	1.098	1.134

^a B = moindres carrés (enfants exclus de la pondération) et C = moindres carrés (enfants inclus dans la pondération).

fondé sur les personnes en utilisant la même information supplémentaire. La seconde série comprenait aussi les enfants répartis en six groupes d'âge-sexe et a servi uniquement à la pondération par les moindres carrés puisque la pondération des enfants par la stratification a posteriori régulière n'a aucun effet sur la pondération des personnes de 15 ans et plus. Bien que tous les estimateurs étudiés soient approximativement non biaisés pour les estimations des caractéristiques de personnes, les hypothèses concernant la nature du sous-dénombrement et de la non-réponse varient de l'un à l'autre (la méthode de compensation de la non-réponse utilisée dans l'enquête sur la population active suppose que les ménages non répondants sont absents de façon aléatoire à l'intérieur d'une région géographique). L'estimateur par stratification a posteriori suppose implicitement que les différences de taux de non-réponse ou de sous-dénombrement dépendent uniquement de l'âge et du sexe et que, par conséquent, elles peuvent être éliminées au moyen d'une estimation fondée sur les personnes et appuyée de l'information supplémentaire portant sur ces caractéristiques. Dans la pondération par les moindres carrés, le poids d'une personne dépendra de la composition âge-sexe du ménage (sans enfant dans un cas et avec des enfants dans l'autre). Ainsi, toutes choses étant égales par ailleurs, la correction appliquée au poids initial d'une personne appartenant à un groupe d'âge-sexe exosé à un fort taux de sous-dénombrement devrait être plus élevée si cette personne demeure seule que si elle demeure avec des personnes qui appartiennent à des groupes d'âge-sexe bien représentés par l'échantillon.

Tableau 1

Nombre de personnes occupées et en chômage, nombre de familles économiques et de personnes seules, enquête sur la population active, mai 1981 (en milliers)

Personnes seules	Familles économiques	En chômage	Occupées	Estimateur ^a
------------------	----------------------	------------	----------	-------------------------

2,432	6,424	850	11,094	A	Canada
2,442	6,446	850	11,090	B	
2,495	6,410	851	11,120	C	
156	563	102	819	A	Région de l'Atlantique
154	570	102	819	B	
156	569	102	821	C	
587	1,723	304	2,725	A	Québec
596	1,725	304	2,724	B	
614	1,714	305	2,735	C	
863	2,325	274	4,198	A	Ontario
861	2,325	273	4,200	B	
881	2,310	273	4,211	C	
506	1,078	83	2,074	A	Région des Prairies
510	1,089	84	2,072	B	
517	1,085	83	2,074	C	
319	735	88	1,277	A	Colombie-Britannique
321	738	88	1,276	B	
327	734	88	1,280	C	

^a A = stratification a posteriori/personne principale, B = moindres carrés (enfants exclus de la pondération) et C = moindres carrés (enfants inclus dans la pondération).

Comme tous les membres du ménage se rattachent au même vecteur ligne de Z et que la probabilité de sélection du premier ordre est la même pour tous, ils auront tous le même poids. De plus, lorsqu'on se sert du poids du ménage pour les personnes, on obtient des résultats compatibles avec les chiffres de population supplémentaires. Bien que cette méthode puisse produire des poids négatifs (si la valeur de quelques-uns des c_j s est inférieure à zéro), les ménages dont le poids est modifié le plus par cette méthode sont en règle générale des ménages dont la composition est peu commune et que l'on retrouve rarement dans l'échantillon et exposés à un taux de non-réponse ou de sous-dénombrement élevé. La méthode proposée a servi récemment à pondérer des données de l'enquête sur la population active portant sur une période de 24 mois; à cette occasion, un seul ménage a reçu un poids négatif, faible par surcroît. Les poids négatifs font problème parce qu'il est difficile de leur attacher le sens que l'on attribue normalement aux poids, c'est-à-dire nombre de personnes ou de ménages dans la population en général représentés par une personne ou un ménage particuliers échantillonnés. Cependant, selon la formule décrite ci-dessus, les poids finals ne sont définis qu'impartialement et on pourrait de fait considérer que ces poids ne sont qu'un moyen commode de produire des estimations. En pratique, il est peu probable qu'une estimation significative de la valeur d'une caractéristique d'intérêt devienne négative sous l'effet de quelques poids négatifs. En revanche, c'est tout autre chose de vouloir faire comprendre à un utilisateur perpexe la notion de poids négatif.

La variance de l'estimateur $\hat{y} = C'x$ défini dans le présent article peut être calculée au moyen des méthodes décrites dans Fuller (1975). De plus, il est possible de montrer que cet estimateur équivaut aux estimateurs de la méthode MCG proposée par Zieschang (1986) et Alexander (1987) lorsque l'espace délimité par les variables auxiliaires Z renferme un vecteur de chiffres 1. Wright (1983) décrit d'autres propriétés de ce genre d'estimateur.

4. RÉSULTATS EMPIRIQUES

L'enquête sur la population active du Canada est une enquête mensuelle à échantillon avec renouvellement qui s'adresse à quelque 48,000 ménages répartis dans tout le pays (voir Platak et Singh 1976 et Singh, Drew et Choudhry 1984). Une fois échantillonnés, les ménages géographiques primaires sont les dix provinces. La taille des échantillons peut varier de 1,500 ménages, à l'Île-du-Prince-Édouard (la plus petite province), à environ 9,000 en Ontario (province la plus populeuse). L'enquête permet de recueillir des données sur la situation des répondants vis-à-vis de l'activité pendant une semaine de référence donnée à chaque mois et aboutit à la publication de toute une série d'estimations relatives au marché du travail dans le pays. Les données d'une des enquêtes mensuelles ont servi à faire une évaluation préliminaire de l'estimateur proposé. Nous avons choisi à cette fin l'enquête de mai 1981 pour pouvoir comparer les résultats à ceux du recensement de 1981 tenu à peu près à cette date. Bien que nous ayons utilisé jusqu'ici les termes "ménage" et "famille" indistinctement, l'utilisateur s'intéresse plus souvent aux estimations touchant la "famille économique", laquelle est constituée de tous les membres d'un ménage qui sont apparentés par le sang, par alliance ou par adoption. Or, l'unité échantillonnée (en l'occurrence, le ménage) se prête mieux théoriquement à la pondération. Néanmoins, les résultats empiriques que nous exposons ici reposent sur des estimations qui ont trait aux familles économiques. Dans cette évaluation, nous sommes intéressés aussi bien aux caractéristiques des personnes (situation vis-à-vis de l'activité) qu'à celles des familles (nombre de familles économiques et de personnes seules). La pondération par les moindres carrés a porté sur deux séries de groupes d'âge-sexe formés par intervalle de cinq ans, les personnes de 70 ans et plus étant groupées selon le sexe. Les enfants de 0 à 14 ans ont été exclus de la première série de groupes (24 au total) pour que l'on puisse comparer l'estimateur proposé à un estimateur régulier de stratification a posteriori

Il est possible de montrer que B sera approximativement non biaisé pour de grands échantillons. Or, le paramètre d'intérêt n'est pas B mais le chiffre de population y . Si nous posons $y = B'x$, y sera un estimateur approximativement non biaisé de y à la condition que $B'x = y$, ou que la somme des résidus pour le modèle de population $Y = XB + E$ soit égale à 0. Cette condition sera respectée si le N -vecteur formé de chiffres 1 est dans l'espace délimité par les colonnes de X et, en particulier, si les variables auxiliaires X renferment un ensemble exhaustif de variables indicatrices qui s'excluent mutuellement (pour les groupes d'âge-sexe par exemple).

Si nous écrivons $y = B'x = Y' \Pi^{-1} T X (X' \Pi^{-1} T X)^{-1} x$, nous constatons que l'estimateur définit implicitement un N -vecteur de poids donné par

$$W = \Pi^{-1} T X (X' \Pi^{-1} T X)^{-1} x,$$

qui ne dépend pas de la variable cible faisant l'objet de l'estimation. Si nous utilisons les poids pour produire des estimations pour les variables auxiliaires, nous obtenons $X'W = x$. Les poids donnent en effet les totaux de population espérés. De plus, si X est constituée entièrement d'un ensemble exhaustif de variables indicatrices qui s'excluent mutuellement, l'estimateur de régression y équivalendra à l'estimateur régulier de la stratification a posteriori. Pour plus de renseignements, voir Bethlehem et Keller (1987).

Il peut être utile de souligner que selon cette méthode, le poids d'une personne quelconque échantillonnée i peut être défini en règle générale de la façon suivante:

$$W_i = \sum^j \frac{\pi_i}{x_i d_i}, \tag{3.3}$$

où $(b_1, \dots, b_p) = (X' \Pi^{-1} T X)^{-1} x$ et π_i est la probabilité de sélection de la personne i . Cela donne à penser que l'on peut modifier la méthode d'estimation décrite ci-dessus de manière à obtenir les poids voulus en définissant les variables auxiliaires de la même façon pour tous les membres du ménage. Une manière simple de le faire est de définir des variables auxiliaires pour le ménage en remplaçant, par exemple, les variables correspondantes pour les personnes par la moyenne du ménage. De façon plus formelle, soit Z une matrice $N \times p$ définie pour une personne i ($i = 1, \dots, N$) appartenant au ménage h ($h = 1, \dots, H$) par

$$Z_{ij} = \frac{n_h}{U_{hj}},$$

où U_{hj} est le total pour la caractéristique j dans le ménage h , c'est-à-dire $U_{hj} = \sum_k X_{kij}$, où la sommation porte sur tous les membres k du ménage h , $n_h =$ taille du ménage h , et $\sum_h n_h = N$. Définissons Y comme un N -vecteur de valeurs d'une variable cible arbitraire définie pour les *personnes*. Comme pour l'estimation relative aux personnes, nous utilisons le modèle de population $Y = ZC + E$ et appliquons les moindres carrés aux données de l'échantillon pour obtenir une estimation

$$\hat{C} = (Z' \Pi^{-1} T Z)^{-1} Z' \Pi^{-1} T Y. \tag{3.4}$$

Nous posons $y = \hat{C}'x$, où x est encore le vecteur des chiffres de population pour les variables auxiliaires. y sera un estimateur approximativement non biaisé de y pourvu que le N -vecteur formé de chiffres 1 se trouve dans l'espace délimité par les colonnes de Z . Comme pour (3.3), le poids d'une personne quelconque échantillonnée dans le ménage h sera défini par

$$W_h = \sum^j \frac{\pi_h n_h}{U_{hj} c_j}. \tag{3.5}$$

moins exposés à la non-réponse et au sous-dénombrement que les ménages de taille plus élevée pourrait corriger en partie cette faiblesse de l'échantillon.

Faute de pouvoir incorporer des données supplémentaires sur les ménages ou les familles dans une méthode de pondération appropriée afin d'attribuer un poids bien défini aux familles, beaucoup de méthodes actuellement en usage attribuent à la famille le poids de la "personne principale" de cette famille. Dans l'enquête sur la population active du Canada, la personne principale est le conjoint de sexe féminin, si elle est présente, sinon c'est le chef de ménage. Comme ces méthodes supposent que les ménages non-dénombés sont absents de façon aléatoire, les estimations produites à l'aide du poids de la personne principale tendent à surestimer les familles de taille plus élevée et à sous-estimer les personnes seules. En outre, on peut estimer de nombreuses caractéristiques (par exemple, population, revenu) à l'aide du poids attribué aux personnes ou du poids attribué aux familles et en règle générale, les deux séries d'estimations différeront entre elles, parfois de façon substantielle. Il est vrai que même dans des conditions d'échantillonnage et d'interview idéales, avec un taux de non-réponse ou de sous-dénombrement uniforme, il y aura toujours des écarts entre les estimations fondées sur les familles et celles fondées sur les personnes pour une même caractéristique. Toutefois, si l'échantillon est suffisamment grand, les différences devraient être faibles. Dans les conditions réelles d'enquête, qui ne sont pas des conditions idéales, les différences sont parfois trop fortes pour que l'on puisse les expliquer simplement par la variabilité d'échantillonnage. En appliquant une méthode d'estimation qui attribue un poids unique au ménage, lequel poids concorde avec les chiffres de population supplémentaires lorsqu'il est utilisé comme poids de personne, nous nous trouvons à résoudre le problème des deux systèmes d'estimation parallèles. C'est ce à quoi nous nous attachons dans la section suivante en définissant un estimateur particulier.

3. ESTIMATEUR PROPOSÉ

Nous allons commencer par définir une méthode de pondération généralisée fondée sur des modèles linéaires de Bethlehem et Keller (1987) et, comme eux, nous allons l'appliquer tout d'abord à l'estimation fondée sur les personnes. Nous modifions ensuite la méthode de manière qu'elle produise des poids de ménages qui conviennent à l'estimation des caractéristiques de personnes. Nous allons ici nous inspirer largement de l'article de Bethlehem et Keller.

Supposons une population cible d'enquête formée de N unités, un N -vecteur Y contenant les valeurs d'une variable cible et une matrice X de dimension $N \times p$ contenant les variables auxiliaires définies pour chaque unité de la population cible. On suppose que les chiffres de population pour chaque variable auxiliaire sont connus et on les désigne collectivement par le p -vecteur x . Dans notre modèle, x sera constitué des totaux de groupes d'âge-sexe. S'il y a corrélation entre les variables auxiliaires et la variable cible, les valeurs de $E = Y - XB$ varieront moins que les valeurs de la variable cible Y pour un p -vecteur B approprié. En appliquant les moindres carrés ordinaires à toutes les unités de la population cible, on obtient

$$B = (X'X)^{-1}X'Y, \quad (3.1)$$

pourvu que X soit une matrice à rang complet. Une estimation de B pour l'échantillon est définie par

$$b = (X'\Pi - {}^1TX)^{-1}X'\Pi - {}^1TY, \quad (3.2)$$

où T est une matrice diagonale dont le i -ième élément est égal à 1 si la i -ième unité de la population est incluse dans l'échantillon, et est égal à 0 dans le cas contraire, et $E(T) = \Pi$.

On a déjà proposé des méthodes qui visaient à déterminer un poids unique pour chaque ménage; on insistait alors sur l'utilisation de données supplémentaires sur les personnes pour améliorer les estimations relatives aux familles. Oh et Scheuren (1978) ont proposé une technique d'estimation itérative "multidimensionnelle" fondée sur la méthode du quotient, qui consiste à corriger successivement par la méthode du quotient les estimations de population des strates formées a posteriori en utilisant les facteurs de correction calculés pour chaque strate, puis à répéter l'opération jusqu'à ce qu'il y ait convergence. Les corrections calculées à chaque étape sont appliquées aux ménages qui comptent des membres dans la strate (formée a posteriori) qui fait l'objet de la correction. Zieschang (1986) a proposé une méthode des moindres carrés généralisés (MCG) selon laquelle on minimise la somme des carrés pondérés des corrections apportées aux poids du plan de sondage (ou poids initiaux) en respectant une série de contraintes linéaires. Alexander (1987) analyse plusieurs méthodes de pondération fondées sur la minimisation d'une fonction de distance, sujette à certaines contraintes. Y compris la méthode MCG, et les évalue par rapport au sous-dénombrement dans les enquêtes. Bien que les méthodes ci-dessus aient été proposées initialement dans le but d'améliorer les estimations relatives aux familles, les poids découlant des divers estimateurs peuvent très bien servir à estimer les caractéristiques de personnes. Cet article préconise l'application d'une méthode d'estimation adaptée aussi bien aux personnes qu'aux familles. Dans la section 2 nous analysons les faiblesses des méthodes d'estimation actuellement en usage pour ce qui a trait aux caractéristiques des personnes et des familles. La section suivante sert à définir un estimateur fondé sur un modèle et inspire d'une méthode de pondération généralisée mise de l'avant par Bethlehem et Keller (1987). La section 4 présente des résultats empiriques tirés de l'enquête sur la population active du Canada. Enfin, la section 5 expose des projets de recherche supplémentaires.

2. MÉTHODES D'ESTIMATION COURANTES

La plupart des enquêtes sur les ménages ont pour objectif principal de produire des estimations pour les caractéristiques de personnes, notamment les caractéristiques de l'activité. S l'unité d'échantillonnage ultime de ces enquêtes est le ménage, c'est surtout pour des raisons d'économie et de commodité. Bien que les premières étapes de la pondération (compensation de la non-réponse, correction en fonction des régions rurales et urbaines, etc.) tiennent compte du ménage en tant qu'unité, ce fait n'est pas reconnu dans l'étape finale (c'est-à-dire que l'on ne tient pas compte du fait que les membres d'un ménage sont échantillonnés indépendamment). De façon plus particulière, les biais de couverture qui pourraient être liés à l'unité échantillonnée ne sont pas directement pris en considération ou compensés dans l'estimation. On suppose donc que le sous-dénombrement est "ignorable" au sens de Rubin (1976); autrement dit, l'étape de l'estimation traite de façon identique toutes les personnes comprises dans une strate formée a posteriori selon les critères d'âge et de sexe, peu importe que ces personnes vivent seules ou qu'elles appartiennent à un ménage formé de plusieurs personnes. Toutefois, une étude sur la non-réponse dans l'enquête sur la population active (Paul et Lawes 1983) a démontré que les ménages de faible taille, particulièrement les ménages sans enfant, tendent à être sous-représentés dans l'échantillon. Bien qu'il n'existe pas d'étude comparable pour les ménages oubliés dans l'enquête sur la population active, des études portant tant sur le sous-dénombrement des ménages privés dans le recensement ont montré que les ménages non-dénombés sont effectivement de plus petite taille en moyenne que les ménages dénombrés (Gosselin et Théroux 1980). Une méthode qui suppose que les ménages sont absents de façon aléatoire peut introduire un biais dans les estimations de la population active pour les personnes, surtout si la répartition des membres des ménages de faible taille par rapport aux caractéristiques de l'activité est différente de celle des membres des ménages de taille plus élevée, toutes choses étant égales par ailleurs. Intuitivement, une méthode d'estimation qui tiendrait compte (ne serait-ce qu'indirectement) du fait que les ménages de faible taille sont

Une méthode intégrée de pondération des personnes et des familles

G. LEMAÎTRE et J. DUFOUR¹

RÉSUMÉ

Les enquêtes sur les ménages utilisent habituellement des méthodes d'estimation distinctes pour les caractéristiques des personnes et celles des familles. Les auteurs proposent dans cet article une méthode intégrée de pondération et définissent à cette fin un estimateur par les moindres carrés. Ils montrent que cet estimateur est sans biais à certaines conditions générales. Au moyen de données de l'enquête sur la population active du Canada, ils calculent les variances de cet estimateur et montrent qu'elles se comparent avantageusement aux variances calculées pour les méthodes usuelles.

MOTS CLÉS: Estimation pour les familles; pondération par les moindres carrés.

1. INTRODUCTION

De nombreuses enquêtes sur les ménages comportent souvent une étape de stratification à posteriori par laquelle on contraindrait les estimations de population de l'enquête, normale-ment classées selon l'âge et le sexe, à respecter des totaux supplémentaires tirés de sources démographiques. Pour faciliter la totalisation, on attribue habituellement à chaque répondant un poids qui est égal au produit de l'inverse du taux de sondage par un facteur de compensation de la non-réponse par un facteur de correction d'âge et le sexe sous forme de quotient. On calcule ensuite l'estimation pour une caractéristique donnée en additionnant les poids attribués à tous les répondants de l'échantillon qui présentent cette caractéristique. En règle générale, les membres d'un même ménage n'ont pas le même poids. Lorsqu'il s'agit d'estimer des caractéristiques de personnes, cela ne pose pas de problème particulier; en revanche, lorsqu'il s'agit de produire des estimations relatives aux ménages ou aux familles, il n'est pas évident lequel de ces poids convient d'utiliser dans les circonstances, s'il y a lieu effective-

ment d'en choisir un. Pour estimer les caractéristiques de familles, on pourrait recourir à l'estimation par le quotient en se servant de données supplémentaires sur les familles et les personnes. Cependant, on dispose rarement de données supplémentaires fiables et récentes sur les familles. À cause d'événements tels qu'une naissance, un décès, un mariage, un divorce ou le départ d'un membre du ménage ou l'arrivée d'un nouveau membre, des caractéristiques comme la taille de la famille varient d'un recensement à l'autre de façon moins prévisible qu'une caractéristique telle que l'âge. Les dossiers administratifs qui nous renseignent sur les variations postcensitaires de la population (c'est-à-dire, registres des naissances, des décès et des migrations) ne disent rien sur les variations touchant les ménages. Le registre des naissances, par exemple, n'indique pas la taille des familles où un enfant est né. Les dossiers fiscaux peuvent combler en partie cette lacune (voir Auger 1987) mais ils ne s'étendent pas à toute la population et ne renferment pas des données suffisamment récentes pour produire des estimations courantes. En l'absence de données supplémentaires sur les familles, on s'est mis à utiliser les poids attribués aux personnes pour estimer les caractéristiques de familles. Pour diverses raisons, cette méthode n'est pas la solution idéale. Dans cet article, nous proposons une méthode d'estimation qui permet de calculer un poids unique pour le ménage, lequel poids peut servir aussi bien à l'estimation pour les personnes qu'à l'estimation pour les familles.

¹ G. Lemaître et J. Dufour, Division des méthodes d'enquêtes sociales, Statistique Canada, 4-ème étage, Immeuble Jean Talon, Parc Tunney, Ottawa (Ontario) K1A 0T6.

- UGH, R.E., TYLER, B.S., et GEORGE, S. (1976). Computer-based procedure for N-dimensional adjustment of data - NJUST. Staff Paper No. 24, U.S. Social Security Administration.
- CHEUREN, F.J. (1981). Methods of estimation for the 1973 exact match study. Dans *Studies from Interagency Data Linkages, Report No. 10*, U.S. Department of Health and Human Services, U.S. Social Security Administration, 9-122.
- TEPHAN, F.F. (1942). An iterative method of adjusting sample frequency tables when expected marginal totals are known. *Annals of Mathematical Statistics*, 13, 166-178.
- VOLTER, K.M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, 81, 338-346.
- WIESCHANG, K.D. (1986a). Generalized least squares: an alternative to principal person weighting. *Population Controls in Weighting Sample Units*, Section 2, Washington, D.C.: U.S. Bureau of Labor Statistics, 1-41.
- WIESCHANG, K.D. (1986b). A generalized least squares weighting system for the Consumer Expenditure Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 64-71.

BIBLIOGRAPHIE

- ALEXANDER, C.H. (1986). The present Consumer Expenditure Surveys weighting method. In *Population Controls in Weighting Sample Units*, Section 7, Washington D.C.: U.S. Bureau of Labor Statistics, 1-32.
- ALEXANDER, C.H., et ROEBUCK, M.J. (1986). Comparison of alternative methods for household estimation. *Proceedings of the Section on Survey Research, American Statistical Association*, 54-64.
- ARORA, H.R., et BRACKSTONE, G.J. (1977). An investigation of the properties of raking ratio estimates: II. With cluster sampling. *Survey Methodology*, 4, 232-252.
- BANKIER, M.D. (1978). An estimate of the efficiency of raking ratio estimators under sample random sampling. *Survey Methodology*, 4, 115-124.
- BRACKSTONE, G.J., et RAO, J.N.K. (1979). An investigation of raking ratio estimators. *Sankhyā*, Série C, 41, 97-114.
- BISHOP, Y.M.M., FIENBERG, S.W., et HOLLAND, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, Massachusetts: MIT Press.
- CRESSIE, N., et READ, T.R.C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society, Série B*, 46, 440-464.
- DARROCH, J.N., et RATCLIFF, D. (1972). Generalized iterative scaling for log-linear models. *Annals of Mathematical Statistics*, 63, 1470-1480.
- DAS GUPTA, P., GIBSON, C., HERRIOT, R.A., LAMAS, E., et ZITTER M. (1986). New approaches to estimating households and their characteristics for states and counties. Article présenté au congrès annuel du Population Association of America de 1986.
- DEMING, W.E., et STEPHAN, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected margins are known. *Annals of Mathematical Statistics*, 11, 427-444.
- FAGAN, J.T., et GREENBERG, B. (1985). Algorithms for making tables additive: raking, maximum likelihood, and minimum chi-square. Statistical Research Division Report Series No. Census/SRD/RR-85/12, U.S. Bureau of the Census.
- FAN, M.C., WOLTMAN, H.F., MISKURA, S.M., et THOMPSON, J.H. (1981). 1980 census variance estimation procedure. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 176-181.
- FIENBERG, S.E. (1986). Comments on some estimation problems in the Consumer Expenditure Surveys. *Dans Population Controls in Weighting Sample Units*, Section 5, Washington, D.C.: U.S. Bureau of Labor Statistics, 1-12.
- HABER, M., et BROWN, M.B. (1986). Maximum likelihood methods for log-linear models when expected frequencies are subject to linear constraints. *Journal of the American Statistical Association*, 81, 477-482.
- HUANG, E.T., et FULLER, W. (1978). Nonnegative regression estimation for sample survey data. *American Statistical Association Proceedings of Social Statistics Section*, 300-305.
- IRELAND, C.T., et SCHEUREN, F.J. (1975). The rakes progress. *Dans Computer Programs for Contingency Table Analysis*, Washington, D.C.: The George Washington University, 155-216.
- KREWSKI, D., et RAO, J.N.K. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics* 9, 1010-1019.
- LUERY, D.M. (1986). Weighting sample survey data under linear constraints on the weights. *Proceedings of the Social Statistics Section, American Statistical Association*, 325-330.
- OH, H.T., et SCHEUREN, F.J. (1978a). Multivariate raking ratio estimation in the 1973 Exact Match Study. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 716-722.
- OH, H.T., et SCHEUREN, F.J. (1978b). Some unresolved application issues in raking ratio estimation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 723-725.

(1986b) montre comment apporter des corrections similaires à l'aide de la méthode des moindres carrés généralisés. Les chiffres de population de ménages réunissent sans aucun doute toutes les caractéristiques nécessaires pour tenir compte des variations de taux de couverture des diverses catégories de ménages. Ils peuvent néanmoins poser des problèmes en ce qui concerne le sous-dénombrement à l'intérieur des ménages puisque leur application peut fausser le calcul de la taille réelle des ménages, ce qui conduirait à une mauvaise classification des ménages de l'échantillon à l'étape de la post-stratification.

6.2 Etudes relatives à la couverture

On peut évaluer assez bien le taux de couverture des personnes en comparant les estimations initiales de l'enquête (N_j) aux totaux de contrôle N_j . Il est difficile de déterminer dans quelle proportion le sous-dénombrement est attribuable à un sous-dénombrement des ménages ou à un sous-dénombrement des personnes à l'intérieur des ménages. On pourrait obtenir des renseignements supplémentaires en comparant les estimations pondérées initiales des ménages aux chiffres de population de ménages, lorsque ces chiffres seraient disponibles. Pour le moment, on pourrait comparer, pour chaque catégorie de ménages, les estimations d'enquête de 1980 et les chiffres correspondants du recensement de la même année. Cependant, les renseignements supplémentaires ne nous permettent pas de faire parfaitement la distinction entre le sous-dénombrement de ménages et le sous-dénombrement à l'intérieur des ménages; pour cela, nous devons absolument poser de nouvelles hypothèses. Alexander et Roebuck (1986) proposent certaines solutions préliminaires pour ajuster des données d'enquête et de recensement à une série de modèles de dénombrement. Wolter (1986) décrit d'autres façons de construire un modèle de dénombrement.

6.3 Estimation de variances

En ce qui concerne la plupart des méthodes fondées sur la distance minimum conditionnelle, on n'a pas cherché à mettre au point des méthodes d'estimation de la variance des estimateurs pondérés. Seule la méthode itérative du quotient a fait l'objet de telles études; voir Arora et Brackstone (1977), Bankier (1978) et Fan et coll. (1981). Dans tous les cas, des méthodes d'itération pourraient servir à estimer la variance des estimateurs pondérés. Ce genre de méthodes s'est avéré raisonnablement efficace dans des conditions assez générales; voir, par exemple, Krewski et Rao (1981). Il reste à déterminer si ces conditions peuvent s'appliquer aux méthodes fondées sur la distance minimum conditionnelle. Zieschang (1986b) a appliqué les méthodes des moindres carrés généralisés à l'enquête sur les dépenses des consommateurs des États-Unis. Scheuren (1981) décrit une application à grande échelle de la méthode itérative du quotient à la pondération des ménages. Les algorithmes du maximum de vraisemblance (EMV-M et EMV-P) n'ont pas été soumis à des applications de ce genre. S'ils devaient servir à la pondération dans les enquêtes, on pourrait devoir se livrer à des recherches en vue d'accroître leur applicabilité.

REMERCIEMENTS

L'auteur tient à exprimer sa reconnaissance à Michael J. Roebuck pour l'aide qu'il lui a apportée dans cette étude, de même qu'au rédacteur associé et aux arbitres pour leurs commentaires utiles. L'auteur tient également à exprimer sa gratitude à Brenda Kelly pour son assiduité dans la préparation de la copie dactylographiée.

En attendant d'autres recherches sur la couverture des enquêtes et ses effets sur la pondération, quelles recommandations pouvons-nous faire? En ce qui concerne les méthodes fondées sur la distance minimum conditionnelle, MCG-M, IMD-M et EMV-M semblent présenter peu d'intérêt à cause de l'irrégularité de leurs résultats dans le contexte d'un sous-dénombrément uniforme des ménages, ceci en dépit du fait que dans un contexte de dénombrement complet, EMV-M semble reposer sur un modèle plus sensible que EMV-P puisqu'en l'occurrence, la dernière unité d'échantillonnage est le ménage et non la personne. La possibilité de poids négatifs nous amène à nous interroger sur l'utilité de la méthode MCG-P, malgré que certaines applications (voir, par exemple, Zieschang 1986b) comportent très peu de poids négatifs de sorte qu'on pourrait leur substituer des poids positifs sans que cela ait d'effet notable sur les estimations. Restent les méthodes IMD-P et EMV-P. Les résultats observés ne permettent pas vraiment de déterminer laquelle de ces méthodes est préférable. Si l'on s'en tient uniquement au calcul, on est porté à choisir la méthode itérative du quotient (IMD-P). Selon les quelques résultats obtenus avec les algorithmes de la section 3, la convergence est plus lente pour les méthodes EMV que pour les méthodes IMD. En outre, de nombreuses recherches ont été faites dans le but de trouver des moyens d'accroître l'efficacité de la méthode itérative du quotient dans des applications à grande échelle (voir, par exemple, Ireland et Scheuren 1975). Compte tenu des remarques précédentes, la méthode itérative du quotient (IMD-P) est celle qui semble offrir les meilleures perspectives parmi les méthodes fondées sur la distance minimum conditionnelle.

Contrairement à la méthode de la personne principale, les méthodes fondées sur la distance minimum conditionnelle produisent des poids de ménages qui correspondent aux totaux de contrôle pour les personnes. Cependant, il n'est pas du tout démontre que les méthodes fondées sur la distance minimum conditionnelle sont plus efficaces que la méthode de la personne principale en ce qui concerne le redressement en fonction du sous-dénombrément. Le sous-dénombrément est un aspect fondamental du problème de la pondération dans les enquêtes. La méthode de la personne principale est un palliatif du sous-dénombrément, qui repose sur des hypothèses très simplistes concernant la couverture des enquêtes. En revanche, nous avons vu dans la section 4 que les méthodes fondées sur la distance minimum conditionnelle pouvaient être considérées comme des estimateurs "optimums" (c'est-à-dire, estimateurs du maximum de vraisemblance ou l'équivalent asymptotique) dans les modèles où l'on pose comme hypothèse le dénombrement complet. Il s'agit donc de choisir entre une solution optimale appliquée à un faux problème et un palliatif de ce qui peut être ou non le vrai problème. Ces observations indiquent qu'il faut manifestement pousser la recherche.

6. QUELQUES SUJETS DE RECHERCHE

6.1 Totaux de contrôle pour les ménages

S'il existait des estimations supplémentaires du nombre de ménages des diverses catégories, on pourrait appliquer les méthodes courantes de post-stratification aux estimations des ménages. On étudierait actuellement la possibilité d'établir des chiffres de population de ménages selon la taille dans le cadre d'un projet qui vise à mettre à jour les données du recensement de 1980 (Das Gupta et coll. 1986). Les chiffres de population de ménages nous permettraient d'envisager d'une toute autre façon le problème de la pondération des ménages.

Les chiffres de population de ménages ne devraient pas exclure pour autant l'utilisation des chiffres de population de personnes. Les chiffres de population de ménages ne sont pas susceptibles de fournir des renseignements détaillés sur l'âge, l'origine raciale et le sexe des membres des ménages. Se servant d'une estimation du nombre total de ménages, Scheuren (1981) applique la méthode itérative du quotient pour comparer simultanément les estimations à des totaux de contrôle indépendants pour les personnes et les ménages. Zieschang

Tableau 3

Sous-dénombrement à l'intérieur des ménages:
Catégories et poids observés,
avec les poids corrigés obtenus selon les trois méthodes

Poids attribué à la catégorie de ménages (composition apparente)		Poids attribué à la catégorie de ménages (composition réelle)	
Personne principale		Personne principale	
EMV-M	EMV-P	EMV-M	EMV-P
Poids total		Poids total	
Catégorie de ménages		Catégorie de ménages	
F	29,000	F	29,000
H	13,500	H	14,997
FF	8,200	FF	7,368
FH	37,200	FH	38,887
HH	4,500	HH	5,623
FHH	10,800	FHH	10,661
FHH	10,800	FHH	12,605
FFF	0	FFF	0
HHH	0	HHH	0
Total	114,000	Total	117,591
			118,274
			116,000
			117,591
			118,274
			116,000

type H ou HH qui sont dénombrés est augmenté en conséquence. Le nombre total pondéré de ménages obtenu par la méthode de la personne principale est égal au nombre de ménages dans la population.

Dans cet exemple, les méthodes fondées sur la distance minimum conditionnelle suresti-

ment le nombre total de ménages mais pondèrent trop faiblement les ménages qui ne comp-

tent pas de personne de sexe masculin. De façon générale, ces méthodes produisent des poids

trop élevés pour les ménages qui comptent des personnes de sexe masculin.

Il ne faudrait pas en conclure que la méthode de la personne principale est toujours

supérieure aux méthodes fondées sur la distance minimum conditionnelle lorsqu'il y a sous-

dénombrement à l'intérieur des ménages. Suivant d'autres hypothèses sur le dénombrement,

elle pourrait ne pas être aussi efficace. De fait, il existe différentes versions de la méthode

de la personne principale selon les enquêtes où elle est appliquée, chaque enquête ayant ses

propres hypothèses sur le dénombrement. Souignons aussi qu'il est possible de combiner

la méthode de la personne principale avec la méthode itérative du quotient; à ce sujet, voir

Scheuren (1981).

Même dans cet exemple, les poids biaisés qui sont produits par les méthodes fondées sur

la distance minimum conditionnelle pourraient être utiles pour l'estimation de certaines

caractéristiques. Si les ménages où des hommes ne sont pas dénombrés tendent à minimiser

la variable d'intérêt, un poids trop élevé pour ces ménages pourrait néanmoins compenser

le biais de réponse lié au sous-dénombrement à l'intérieur des ménages.

L'exemple le plus frappant de cette utilité est l'estimation du nombre total d'hommes;

dans ce cas, les poids calculés selon EMV-M et EMV-P correspondent aux totaux de con-

trôle, ce qui n'est pas le cas des poids calculés selon la méthode de la personne principale.

Toutefois, les poids biaisés ne sont pas souhaitables lorsqu'il s'agit de caractéristiques qui

font peu souvent l'objet d'erreurs de déclaration attribuables au sous-dénombrement de per-

sonnes de sexe masculin (par exemple, mode d'occupation: locataire/ propriétaire). L'efficacité

des méthodes de pondération dans des cas comme ceux exposés ici dépend incontestable-

ment de la nature du sous-dénombrement et de sa relation avec la variable estimée. Alex-

ander et Roebuck (1986) traitent plus en détail le sujet et fournissent d'autres exemples.

Tableau 2

Sous-dénombrement à l'intérieur des ménages:
Description de la population et de l'échantillon

Catégorie de ménages (composition réelle)	Catégorie de ménages (composition apparente)	Population réelle	Poids totaux initiaux	
1: F	F	25,000	25,500	0
2: H	H	13,500	13,500	0
3: FF	FF	7,000	7,000	0
4: FH	FH	36,000	36,000	7,000
5: HH	HH	4,500	4,500	4,000
6: FHH	FHH	10,800	10,800	4,500
7: FHH	FHH	10,800	10,800	0
8: FFF	FFF	0	1,200	1,200
9: HHH	HHH	0	0	0
Chiffres de population:			116,000	114,000
			115,000	
			101,000	
Comptes initiaux			Femmes	Hommes
			115,000	90,900

Aucune des colonnes de chiffres du tableau 2 n'est observée puisqu'il n'y a pas de com-
ptage de contrôle de ménages. En outre, on ne connaît pas le mode de composition réel des
ménages auxquels appartiennent les unités de l'échantillon. Ainsi, les ménages de type F[H]
(catégorie 4) passent pour des ménages de type F (catégorie 1). On observe plutôt les poids
totaux initiaux des ménages qui semblent avoir un mode de composition donné. Le tableau
3 donne les poids corrigés calculés selon trois méthodes: EMV-M, EMV-P et personne prin-
cipale. Comme les méthodes MCG-M et IMD-M produisent des résultats comparables à ceux
de EMV-M et que les méthodes MCG-P et IMD-P produisent les mêmes résultats que EMV-
P, nous n'avons pas cru nécessaire de reproduire tous ces résultats.

Les trois dernières colonnes du tableau 3 contiennent les poids totaux corrigés qui ont
été calculés pour chaque catégorie de ménages (composition réelle) à l'aide des méthodes
EMV-M et EMV-P et de la méthode de la personne principale. On constate que les poids
obtenus par la méthode de la personne principale pour chaque catégorie "réelle" correspon-
dent aux chiffres de la population réelle qui figurent dans la troisième colonne du tableau
1. C'est pourquoi on peut dire que la méthode de la personne principale produit des poids
non biaisés.

Cet exemple reprend les hypothèses qui sous-tendent la version simplifiée de la méthode
de la personne principale. Il est donc normal, en l'occurrence, que cette méthode produise
des résultats très satisfaisants. Il va de soi que les ménages de type [H] ou [HH] qui échap-
pent entièrement à l'enquête n'ont pas de poids; en contrepartie, le poids des ménages de

Nous présenterons deux exemples qui reproduisent deux formes de sous-dénombrement contraires. Dans le premier exemple (sous-dénombrement de ménages), nous supposons qu'il y a un taux uniforme de sous-dénombrement des ménages de 10% mais qu'il n'y a pas de sous-dénombrement à l'intérieur des ménages. Dans le second exemple (sous-dénombrement des hommes à l'intérieur des ménages), nous supposons qu'il y a un taux de sous-dénombrement des hommes de 10% attribuable au sous-dénombrement à l'intérieur des ménages qui compte des personnes des deux sexes et au sous-dénombrement des ménages qui ne compte que des personnes de sexe masculin. Pour les ménages formés d'une seule personne, le "sous-dénombrement à l'intérieur des ménages" signifie que le ménage au complet échappe à l'enquête.

Le tableau 1 décrit le premier exemple. On constatera que chaque catégorie de ménages dans l'échantillon est sous-représentée dans une proportion de 10%. Dans le cas d'un échantillon suffisamment grand, cette sous-représentation serait sûrement attribuable au sous-dénombrement systématique plutôt qu'à l'erreur d'échantillonnage. En appliquant les méthodes fondées sur la distance minimum conditionnelle et la méthode de la personne principale, nous obtenons pour chaque catégorie de ménages les poids totaux corrigés qui figurent dans les quatre dernières colonnes du tableau.

On remarquera que les méthodes MCG-P, IMD-P et EMV-P produisent toutes des poids équivalents à la population réelle. Il s'agit donc de poids "non biaisés". Comme toutes les personnes ont un facteur du second degré de 1/9, la méthode de la personne principale produit aussi des poids équivalents à la population réelle. Les autres méthodes, MCG-M, IMD-M et EMV-M, produisent toutes des poids beaucoup trop faibles pour les ménages formés d'une seule personne et des poids beaucoup trop élevés pour les ménages formés de trois personnes. Cela est compréhensible intuitivement. Puisque ces trois méthodes n'admettent pas le sous-dénombrement systématique et qu'elles doivent expliciter la sous-représentation de personnes par l'erreur d'échantillonnage, l'application la plus plausible est que l'échantillon comportait par hasard un nombre de ménages de trois personnes inférieur à la moyenne. Le rendement supérieur de EMV-P est compréhensible puisque cette méthode repose sur un modèle d'échantillonnage multinomial, selon lequel les personnes sont échantillonnées sans égard aux ménages.

Au point de vue pratique, cet exemple déprécie largement les méthodes MCG-M, IMD-M et EMV-M. Un taux de sous-dénombrement uniforme n'empêchera pas ces méthodes de fausser la distribution de la taille des ménages. Chose encore plus grave toutefois, la distortion créée par l'application de ces méthodes dans l'exemple est contraire au fait généralément admis en ce qui concerne les différences de taux de dénombrement des ménages, c'est-à-dire que les ménages de faible taille sont plus susceptibles d'échapper à l'enquête que les ménages de grande taille et que, par conséquent, ils devraient avoir des poids relativement plus élevés et non des poids relativement moins élevés.

Le deuxième exemple met l'accent sur le sous-dénombrement des hommes à l'intérieur des ménages. Le cas est plus complexe que dans l'exemple précédent parce que la composition apparente d'un ménage peut être différente de sa composition réelle. Par exemple, un ménage composé d'un homme et d'une femme peut sembler être un ménage d'une seule personne. Nous allons différencier la composition réelle de la composition apparente en modifiant la notation utilisée jusqu'à maintenant. Par exemple, un ménage de type FH où l'homme n'est pas dénombré sera désigné par F[H]. Un ménage désigné par [H] ou [HH] aura tout simplement été omis. Le tableau 2 contient les données hypothétiques. La population réelle est la même que celle de l'exemple précédent.

On observe un taux de sous-dénombrement des hommes de 10% à cause du sous-dénombrement à l'intérieur des ménages et du sous-dénombrement des ménages composés uniquement de personnes de sexe masculin. Chaque homme a 10% de chances d'échapper à l'enquête.

Tableau 1
Sous-dénombrement des ménages:
Description de la population et de l'échantillon

Poids total (W_i) pour les méthodes:

Personne principale	MCG-P	IMD-P	EMV-P
1: F	25,000	25,000	23,704
2: H	15,000	15,000	14,075
3: FF	7,000	7,000	7,013
4: FH	40,000	39,632	39,672
5: HH	5,000	4,900	4,906
6: FFH	12,000	12,594	12,506
7: FHH	12,000	12,449	12,428
8: FFF	0	0	0
9: HHH	0	0	0
Total	116,000	114,367	114,370
Chiffres de population:			
	104,400	114,483	115,000
	0	0	0
	10,800	12,408	12,428
	4,500	4,913	4,906
	36,000	39,708	39,672
	6,300	7,020	7,016
	14,120	14,097	14,075
	22,500	23,785	23,745
Hommes			
	103,500	103,500	103,500
Femmes			
	101,000	101,000	101,000
Comptes initiaux pondérés:			
	116,000	114,367	114,370

catégories de ménages, le nombre de catégories incluses dans un échantillon ne peut jamais dépasser la taille de cet échantillon et est, le plus souvent, beaucoup moindre. C'est ce qu'Ireland et Scheuren (1975) ont constaté pour des cellules de ménages connexes. Le simple fait de réduire le volume des calculs en combinant les poids des ménages à personne unique à une même catégorie peut avoir des avantages; c'est ce qu'a fait le U.S. Bureau of Labor Statistics en appliquant la méthode des moindres carrés généralisés à l'enquête sur les dépenses des consommateurs.

La version simplifiée de la méthode de la personne principale dépend aussi uniquement du mode de composition du ménage. Si deux ménages ont le même mode de composition, ceux personnes principales se trouveront dans la même cellule de post-stratification, celle du facteur de stratification a posteriori est le plus près de 1. On utiliserait donc le même facteur de correction pour les deux ménages. Selon la version fondamentale de la méthode de la personne principale, la personne principale est en partie définie en fonction de celui ou celle qui se trouve être la personne repère. Dans ce cas, le facteur de correction n'est donc pas entièrement déterminé par le mode de composition du ménage.

Il convient de souligner que la méthode EMV-M équivaut à calculer des estimations multinomiales du maximum de vraisemblance (moyennant le respect de la condition (1)) de $p_i, i = 1, \dots, T$, où p_i est la proportion de ménages de la catégorie i dans la population. La méthode EMV-P peut être interprétée de la même manière. Aucun de ces modèles, qui se rattachent aussi aux méthodes MCG et IMD correspondantes, n'admet le sous-dénombrement systématique.

5. ANALYSE DES METHODES

Nous allons tout d'abord formuler des hypothèses sur les propriétés des méthodes fondées sur la distance minimum conditionnelle en nous servant des résultats de la section 4 et nous allons poursuivre avec des exemples fictifs simples qui, de façon générale, semblent appuyer es hypothèses.

La première hypothèse est que les méthodes EMV-M, MCG-M et IMD-M tendront à produire des résultats similaires et que les méthodes EMV-P, MCG-P et IMD-P tendront à se rapprocher l'une de l'autre, du moins pour de grands échantillons. Cette hypothèse repose sur le fait que les estimateurs de ces méthodes sont tous de meilleurs estimateurs asymptotiques normaux selon le modèle d'échantillonnage multinomial pertinent, où les cellules représentent les catégories de ménages. En ce qui concerne les échantillons de taille faible ou moyenne, on peut s'attendre à des écarts plus marqués entre les résultats des diverses méthodes, surtout s'il y a un grand nombre de modes de composition des ménages, ce qui crée des échantillons de faible taille dans les "cellules" du modèle multinomial.

Les exemples présentés plus loin tendent à confirmer cette hypothèse; les méthodes pour les ménages, tout comme celles pour les personnes, produisent toutes des résultats très comparables, même lorsque les données hypothétiques s'ajustent mal au modèle. Toutefois, comme ces exemples ne concernent qu'un petit nombre de catégories de ménages et de cellules de post-stratification, ils servent plus à illustrer l'hypothèse qu'à la vérifier.

La seconde hypothèse repose sur une analyse qui vise à déterminer le genre de modèles d'échantillonnage où l'on peut appliquer plus particulièrement les fonctions estimatrices du maximum de vraisemblance ou des approximations asymptotiques de celles-ci. Dans ce genre de modèles, on suppose que le dénombrement est complet. On suppose aussi une distribution correspondant à des probabilités qui représentent les proportions réelles dans la population et qui sont conformes aux totaux de contrôle "réels" utilisés dans l'équation (1). Selon ce genre de modèles, les estimations initiales de l'échantillon se rapprocheraient des totaux

4. RÔLE DU "MODE DE COMPOSITION" D'UN MÉNAGE

En ce qui concerne les six méthodes fondées sur la notion de distance minimum conditionnelle, le rapport du poids initial d'un ménage à son poids corrigé dépend de la répartition des membres de ce ménage dans les diverses cellules de post-stratification. Avant de poursuivre cette analyse, nous devons définir ici le "mode de composition" d'un ménage. Deux ménages d'un échantillon (par exemple k et m) auront le même mode de composition s'ils comptent exactement le même nombre de personnes dans chacune des cellules de post-stratification, c'est-à-dire si

$$a_{kj} = a_{mj} \text{ pour } j = 1, \dots, J. \quad (8)$$

On pourrait avoir, par exemple, un ménage constitué d'un homme de race blanche âgé de 35 à 39 ans et d'une femme de race blanche âgée de 30 à 34 ans. Il convient de souligner que le mode de composition ne dépend pas des liens de parenté.

Le rapport du poids corrigé au poids initial, W^k/S^k , est identique pour tous les ménages qui ont le même mode de composition. Autrement dit, si k et m satisfont l'équation (8), alors $W^k/S^k = W^m/S^m$. Ireland et Scheuren (1975) ont appliqué cette relation. Alexander et Roebuck (1986) en font la démonstration.

Une conséquence avantageuse de cette relation est que la catégorie de ménages peut remplacer le ménage comme unité d'analyse dans le calcul des poids pour les méthodes fondées sur la distance minimum conditionnelle. Un exemple simple nous permettra de mieux saisir les conséquences de ces observations. Supposons qu'il y a deux cellules de post-stratification: $j = 1$ pour les femmes et $j = 2$ pour les hommes. L'échantillon est formé de K ménages. Pour le ménage k , le vecteur (a_{k1}, a_{k2}) définit le nombre de femmes et d'hommes qui constituent le ménage. Un vecteur $(2, 1)$ signifie que le ménage compte deux personnes de sexe féminin et une de sexe masculin.

En pratique, un ménage ne peut dépasser une certaine taille et le nombre de catégories de ménages est aussi limité. Pour les besoins de la cause, supposons qu'aucun ménage ne compte plus de trois personnes. Nous avons donc $T = 9$ modes de composition de ménage qui correspondent aux vecteurs $(1, 0)$, $(0, 1)$, $(2, 0)$, $(1, 1)$, $(0, 2)$, $(2, 1)$, $(1, 2)$, $(3, 0)$, $(0, 3)$. Ces modes seront désignés dans l'ordre par $t = 1, \dots, 9$. On leur attribuera aussi, dans l'ordre, les codes mnémotechniques $F, H, FF, FH, HH, FFF, FHH, FHF, HHH$. Le tableau 1 contient des données d'échantillon et des totaux de contrôle hypothétiques. Précisons que S est le poids total initial attribué aux ménages de la catégorie t .

Les corrections de poids selon les méthodes fondées sur la distance minimum conditionnelle peuvent être effectuées à l'aide des poids attribués aux modes de composition des ménages S_1, \dots, S_9 sans tenir compte des poids des ménages proprement dits. Nous pouvons calculer les poids corrigés W_1, \dots, W_9 au moyen des algorithmes présentés dans la section 3 en effectuant la sommation en fonction de t plutôt qu'en fonction de k . Ainsi, le poids corrigé d'un ménage de catégorie t est défini comme le produit de W^t/S^t par le poids initial de ce ménage. (Malgré le risque de confusion, nous avons choisi de désigner le poids des ménages par S^t et le poids total d'une catégorie de ménages par S^t pour souligner le fait que les formules définies dans les sections 2 et 3 s'appliquent aussi bien aux catégories de ménages qu'aux ménages proprement dits. Au moment des calculs, le contexte permet de faire la distinction entre S^k et S^t .)

Le fait de pouvoir utiliser des catégories de ménages comme unités d'analyse au lieu des ménages proprement dits est très avantageux pour la présentation d'exemples simples. Même lorsque l'analyse s'étend aux 48 cellules de post-stratification, il peut être encore avantageux d'utiliser les catégories de ménages. Malgré qu'il existe en théorie un nombre indéfini de

1 Solution pour IMD-M

L'équation des poids est

$$(7) \quad W^k = S^k \prod_{j=1}^f \gamma_j a_{kj}$$

condition que (1) soit respectée. S'il est possible de déterminer des valeurs, $\gamma_1, \dots, \gamma_f$ telle sorte que les poids calculés selon (7) satisfassent l'équation (1), alors ces poids minimisent (2b) à condition que (1) soit respectée. Nous décrivons ci-dessous un algorithme d'itération qui produit un vecteur de poids \bar{W} .

Posons $W^k(0) = S^k$ et $\gamma^{(0)} = 1$. Puis, à la i -ème itération, posons

$$\gamma_j(i) = \gamma_j(i-1) \left[1 - (N_j(i) - 1) - N_j(i) / \sum_{s=1}^s a_{sj}^2 W^s(i-1) \right],$$

$$N_j(i-1) = \sum_{s=1}^s a_{kj} W^s(i-1). \text{ Enfin, posons } W^k(i) = S^k \prod_{j=1}^f (\gamma_j(i))^{a_{kj}}.$$

2 Solution pour EMV-M

La solution a la forme

$$W^k = S^k / \left(1 + \sum_{j=1}^f \gamma_j a_{kj} \right).$$

condition que (1) soit respectée.
Une solution itérative est

$$W^k(0) = S^k \quad \text{et} \quad \gamma_f(0) = 0,$$

$$\gamma_j(i) = \gamma_j(i-1) + (N_j(i) - 1) - N_j(i) / \left(\sum_{s=1}^s a_{sj} W^s(i-1) \right)^2 / S^k.$$

$$W^k(i) = S^k / \left(1 + \sum_{j=1}^f \gamma_j(i) a_{kj} \right).$$

3 Solution pour EMV-P

La solution a la forme

$$W^k = S^k / \left(\sum_{j=1}^f \gamma_{kj} a_{kj} / a_k \right),$$

condition que (1) soit respectée.
Une solution itérative est

$$W^k(0) = S^k \quad \text{et} \quad \gamma_f(0) = 1,$$

$$\gamma_j(i) = \gamma_j(i-1) N_j(i) / N_j,$$

$$W^k(i) = S^k / \left(\sum_{j=1}^f \gamma_j(i) a_{kj} / a_k \right).$$

3. CALCUL DES POIDS

Les deux méthodes des moindres carrés, MCG-M et MCG-P, ont des expressions en forme analytique pour \bar{W} , à condition qu'il existe une solution pour l'équation (1). En ce qui concerne la méthode MCG-M, l'équation des poids corrigés est

(6)
$$\bar{W} = \bar{S} + MA(A'MA)^{-1}(\bar{N} - A'\bar{S})$$

où $\bar{S} = (S_1, \dots, S_K, \bar{N}) = (N_1, \dots, N_J, A)$ est la matrice (a_{kj}) et M est la matrice diagonale $K \times K$, dont la diagonale principale est constituée des éléments de \bar{S} . En ce qui a trait à la méthode MCG-P, les poids \bar{W} sont aussi définis par l'équation (6) sauf que M est une matrice diagonale $K \times K$ dont la diagonale principale est constituée des valeurs $S_1/a_1, \dots, S_K/a_K$.

L'inconvénient de la solution (6) pour l'une ou l'autre des méthodes précitées est qu'elle peut produire des poids négatifs, ce qui est déroulant au point de vue théorique et inacceptable au point de vue pratique. Il est habituellement possible d'introduire de nouvelles conditions pour faire en sorte que les poids soient positifs. Zieschang (1986a) et Huang et Fuller (1978) montrent comment. Toutefois, l'introduction de nouvelles conditions élimine l'avantage que présente une expression en forme analytique simple.

La méthode itérative du quotient (IMD-P) a déjà servi à la pondération des ménages (voir, par exemple, Oh et Schuuren (1978a)). Une méthode connexe, qui a été largement expérimentée, est décrite dans Pugh, Tyler et George (1976), ceux-ci s'étant fondés sur l'approche de Stephan (1942). S'inspirant de Darroch et Ratcliff (1972), Luery (1986) définit un algorithme d'itération qui s'avère convergent lorsqu'il existe une solution à l'équation (1). Nous décrivons ici cette méthode étant donné que le processus itératif est facile à interpréter. Le processus débute par les poids à l'étape 0.

$$W^k(0) = S_k(N/N)$$

Autrement dit, le poids initial S_k est multiplié par un indice de correction global qui correspond au quotient de la population connue N , par la population totale pondérée. Pour les itérations suivantes, la correction est définie par

$$W^k(i) = W^k(i-1) \prod_{j=1}^f \left(N_j / \sum_{s=1}^s a_{sj} W^s(i-1) \right)^{a_{kj}/a_k}$$

Vous remarquerez que $W^k(i-1)$ est multiplié par la moyenne géométrique des facteurs de stratification a posteriori qui s'appliquent aux membres du k -ième ménage, ces facteurs étant calculés à l'aide des poids de l'itération $i-1$.

Les trois autres méthodes (IMD-M, EMV-M et EMV-P) ont fait l'objet de peu d'analyses. Les algorithmes d'itération décrits ci-dessous se sont avérés efficaces dans des exemples fictifs simples comme ceux présentés dans la section 5. Dans chaque cas, on peut déterminer à l'aide des multiplicateurs de Lagrange un système d'équations que les poids doivent satisfaire pour minimiser le critère de distance assujéti aux conditions. Il n'est pas possible de résoudre directement ces équations mais s'il se trouve une méthode itérative qui produit des solutions convergentes, les solutions satisfieront les équations. Toutefois, on n'a pu vérifier de façon générale la convergence de ces algorithmes. L'approche de Haber et Brown (1986) pourrait être une solution alternative dans le cas des critères du "maximum de vraisemblance". Fagan et Greenberg (1985) proposent également des solutions à cet égard.

(5a) MCG:
$$\sum^k a_k. (W_k - S_k)^2 / S_k,$$

(5b) IMD-P:
$$\sum^k a_k. S_k - \sum^k a_k. W_k + \sum^k a_k. W_k \ln (W_k / S_k),$$

(5c) EMV-P:
$$\sum^k a_k. W_k - \sum^k a_k. S_k - \sum^k a_k. S_k \ln (W_k / S_k).$$

ces critères sont devenus des sommes en fonction des ménages mais la taille du ménage W_k y a été incluse dans le but de mesurer la distance entre le vecteur de poids initial et le vecteur de poids corrigé. Nous verrons que ces critères présentent des avantages par rapport à la méthode plus directe qui a servi à établir (2a), (2b) et (2c).

3.3 Méthode de la personne principale

Dans la version de base de la méthode de la personne principale, le poids de post-tratification attribué à la "personne principale" du ménage sert de poids pour le ménage. Pour déterminer la personne principale, il faut tout d'abord déterminer la personne repère du ménage. Pour l'interviseur, la personne repère sera celle dont le nom figure à côté de la question suivante: "Donnez-moi tout d'abord le nom du propriétaire ou du locataire du logement".

On définit les rapports entre les membres du ménage en fonction de leur rapport avec la personne repère. Pour les besoins de la cause, la notion de "personne repère" est substituée à celle de "chef de ménage".

Si la personne repère est un homme marié qui vit avec sa femme, celle-ci est reconnue comme la personne principale. Autrement, la personne repère devient la personne principale. Des critères reposent sur l'idée que la personne principale doit être une personne qui est peu susceptible d'échapper à l'enquête s'il y avait sous-dénombrement à l'intérieur du ménage. En règle générale, les femmes échappent moins souvent aux enquêtes que les hommes. En outre, le propriétaire ou le locataire d'un logement peut difficilement échapper à l'enquête. Le principe de base de la méthode de la personne principale veut qu'il n'y ait qu'une personne principale par ménage. Ainsi, l'estimation du nombre de personnes principales suffit pour estimer le nombre de ménages. Cette méthode est utilisée dans la U.S. National Crime Survey. D'autres enquêtes, comme la U.S. Consumer Expenditure Survey ou la Current Population Survey, comportent des mesures de redressement additionnelles fondées sur des hypothèses concernant le sous-dénombrement de personnes principales à l'intérieur des ménages par comparaison au sous-dénombrement d'autres personnes dans la même cellule (Alexander 1986).

Il est difficile de construire un modèle théorique de la méthode de la personne principale parce que la définition de la personne repère est quelque peu arbitraire. Dans les exemples cités de la section 5, nous allons utiliser une version simplifiée de la méthode de la personne principale, selon laquelle la personne principale est le membre du ménage dont la cellule de post-stratification présente le meilleur taux de dénombrement, c'est-à-dire celle dont le facteur de post-stratification est le plus près de un. Scheuren (1981) applique une version semblable. Cette version simplifiée de la méthode de la personne principale sera représentée de la façon suivante. Pour le k -ième ménage de l'échantillon, soit $f(k)$ la cellule de post-stratification laquelle appartient la personne principale de ce ménage. Alors, le poids attribué à cette personne est

$$W_k = S_k (N_{f(k)} / N_{j(k)}).$$

Dans chaque cas, $D(\bar{W}, \bar{S})$ est non négatif et est égal à zéro si et seulement si $\bar{W} = \bar{S}$. On peut le vérifier, de la façon habituelle, en examinant les dérivées partielles du premier et du second ordre de chaque expression par rapport à W_k .

Dans la section 3, nous étudierons des algorithmes permettant de calculer \bar{W} de manière à minimiser ces trois critères tout en respectant la condition (1) au niveau d'approximation voulu.

2.2 Méthodes fondées sur les poids attribués aux personnes

La question peut être envisagée sous un autre angle; cette approche amène toutefois une modification légère mais importante des trois critères de distance. Les critères modifiés sont définis par les expressions (5a), (5b) et (5c) ci-dessous. Bien que ces critères servent à déterminer des poids pour les ménages, ils sont le produit d'une méthode qui, au départ, cherche à définir des poids pour les personnes. En conséquence, nous devons tout d'abord déterminer, pour les personnes, les poids qui se rapprochent le plus possible des poids initiaux des ménages auxquels appartiennent ces personnes, à condition que l'estimation pondérée du nombre de personnes dans chaque cellule de post-stratification égale le chiffre de population. Désignons les membres du k -ième ménage par $i = 1, \dots, a_k$, et définissons S_{ki} comme le poids initial de la i -ième personne dans le k -ième ménage; remarquons que $S_{ki} = S_{kij}$. Soit b_{kij} une variable indicatrice (0-1) qui montre si la i -ième personne du k -ième ménage se trouve dans la j -ième cellule de post-stratification. Alors, la condition pour qu'il y ait correspondance avec les chiffres de population est

$$(3) \quad \sum_{i=1}^k \sum_{j=1}^I b_{kij} W_{ki} = N_j.$$

Les trois critères pour la pondération des personnes seraient

$$(4a) \quad \sum_{i=1}^k \sum_{j=1}^I (W_{ki} - S_{ki})^2 / S_{ki},$$

$$(4b) \quad S_{..} - W_{..} + \sum_{i=1}^k \sum_{j=1}^I W_{ki} \ln(W_{ki} / S_{ki}),$$

$$(4c) \quad W_{..} - S_{..} - \sum_{i=1}^k \sum_{j=1}^I S_{ki} \ln(W_{ki} / S_{ki}).$$

Ces critères peuvent servir à déterminer des poids pour les personnes. De fait, le critère (4c) produit des poids de post-stratification qui servent à la pondération des personnes dans l'enquête sur les dépenses des consommateurs (voir Alexander 1986). Cependant, nous cherchons ici à déterminer des poids pour les ménages. Nous pouvons déterminer de tels poids à l'aide des fonctions de critère ci-dessus si nous posons comme condition additionnelle que tous les membres d'un même ménage aient la même pondération. Par conséquent, posons $W_{ki} = W_k$ pour $i = 1, \dots, a_k$. Sous cette condition, (3) devient

$$N_j = \sum_{k=1}^K \left(\sum_{i=1}^{a_k} b_{kij} \right) W_{ki} = \sum_{k=1}^K a_k W_k,$$

qui est identique à l'équation (1) de la section 2.1. Les critères de distance (4a), (4b) et (4c) s'écrivent maintenant:

Supposons qu'il y a J cellules de post-stratification et que l'on connaît le nombre de personnes (N_j) dans chacune. Par exemple, dans le cas de l'enquête sur les dépenses des consommateurs aux Etats-Unis, il y a $J = 48$ cellules formées par la combinaison des deux origines raciales (noir, non noir) et de douze catégories d'âge. Les personnes de moins de 14 ans ne sont pas visées par cette enquête. Les chiffres de population pour ces cellules sont définis en l'occurrence par un vecteur $\bar{N} = (N_1, \dots, N_J)$.

La composition des ménages de l'échantillon est définie par une matrice $A = (a_{kj})$, où j est égal au nombre de personnes du k -ième ménage qui se trouvent dans la j -ième cellule de post-stratification. La sommation des cellules de post-stratification pour le k -ième ménage donne a_k , soit le nombre total de personnes que compte le ménage. Le vecteur (a_{k1}, \dots, a_{kJ}) décrit la composition du ménage k . Par exemple, un vecteur $(2, 1, 0, 0, \dots, 0)$ signifie que deux personnes du ménage se trouvent dans la première cellule et une troisième dans la seconde.

Etant donné les poids initiaux \bar{S} , l'estimation pondérée du nombre de personnes dans la cellule j sera $N_j = \sum_k a_{kj} S_k$ ou, de façon générale, $\bar{N} = A' \bar{S}$.

En règle générale, $\bar{N} \neq \bar{N}$, autrement dit, l'estimation pondérée initiale du nombre de personnes dans les cellules de post-stratification peut différer de l'effectif réel de la cellule. Nous cherchons ici à définir un nouveau vecteur de poids $\bar{W} = (W_1, \dots, W_K)$ pour les ménages de l'échantillon de sorte que $\bar{N} = A' \bar{W}$ ou

$$(1) \quad \sum^k a_{kj} W_k = N_j \text{ pour } j = 1, \dots, J.$$

Equation (1) n'a pas nécessairement une solution unique. Les méthodes fondées sur la notion de distance minimum conditionnelle consistent essentiellement à choisir un vecteur \bar{W} de manière à minimiser la distance $D(\bar{W}, \bar{S})$ entre les vecteurs \bar{W} et \bar{S} à condition que l'équation (1) soit respectée. Ainsi, en respectant la condition d'équivalence entre les poids corrigés et les totaux de contrôle, nous nous trouvons à modifier le moins possible les poids initiaux \bar{S} . Il convient de souligner que pour certaines valeurs N_1, \dots, N_J , il peut être impossible de calculer un vecteur de poids \bar{W} qui satisfasse à la condition (1). En pratique, cette éventualité ne semble pas poser de problème si l'échantillon est suffisamment grand pour inclure un nombre raisonnable d'unités des diverses catégories de ménages, puisque les totaux de contrôle \bar{N} sont tirés de la population réelle et que par conséquent, il peut les considérer comme "possibles".

Il y a plusieurs façons de mesurer la différence entre deux vecteurs. Nous allons étudier trois critères de distance $D(\bar{W}, \bar{S})$ qui correspondent respectivement, au niveau des ménages, une fonction objective des moindres carrés généralisés (MCG-M), à une fonction d'information minimum discriminante (IMD-M) et à une fonction estimatrice du maximum de vraisemblance (EMV-M). Les critères sont:

$$(2a) \quad \text{MCG - M:} \quad \sum^k (W_k - S_k)^2 / S_k,$$

$$(2b) \quad \text{MDI - H:} \quad (S_j - W_j) + \sum^k W_k \ln(W_k / S_k),$$

$$(2c) \quad \text{MLE - H:} \quad (W_j - S_j) - \sum^k S_k \ln(W_k / S_k).$$

ans cet article, les points au bas des lettres désignent la sommation par rapport à un indice inférieur.

Toutefois, cette méthode a un inconvénient majeur. En effet, lorsqu'on se sert des poids de ménage obtenus par cette méthode pour calculer des estimations pondérées du nombre de personnes dans chaque cellule de post-stratification, chaque personne étant affectée d'un poids du ménage auquel elle appartient, ces estimations ne correspondent pas aux chiffres de population utilisés dans la post-stratification. C'est pourquoi on s'est intéressé à de nouvelles méthodes de pondération des ménages qui produisent des poids à partir desquels on obtient nécessairement des estimations conformes aux chiffres de population. Le présent article vise à exposer une série de méthodes de pondération des ménages qui produisent des poids à partir desquels on obtient nécessairement des estimations conformes aux chiffres de population des diverses cellules. Le principe de base de ces méthodes est d'utiliser les chiffres de population des diverses cellules. Le principe de base de ces méthodes est d'utiliser les chiffres de population des diverses cellules. Le principe de base de ces méthodes est d'utiliser les chiffres de population des diverses cellules.

Dans la section 2, nous décrivons six méthodes de pondération fondées sur la notion de "distance minimum conditionnelle" et une version de la méthode de la personne principale. Trois des six méthodes précitées ont déjà fait l'objet d'analyses; les autres sont présentées pour donner une image plus complète de la question. La section 3 porte sur le calcul des poids. Dans la section 4, nous voyons comment la composition du ménage influe sur la correction des poids. Dans la section 5, nous analysons les résultats et présentons des exemples qui peuvent aider à comprendre le mécanisme de ces méthodes. Enfin dans la section 6, nous proposons des sujets de recherche.

Cet article n'est pas le premier du genre. Luey (1986) propose d'appliquer la série de méthodes fondées sur la notion de distance minimum conditionnelle à la pondération des ménages. Poursuivant le travail de Luey, Zieschang (1986a) propose d'appliquer l'une de ces méthodes, la méthode des moindres carrés généralisés, à la pondération dans les enquêtes sur les dépenses des consommateurs aux États-Unis. Une autre méthode de la série est la "méthode de l'information minimum discriminante", connue aussi sous le nom de "estimation itérative par la méthode du quotient" ou simplement "méthode itérative du quotient". Oh et Scheuren (1978a) traitent particulièrement de l'application de cette méthode à la pondération des ménages et enrichissent la bibliographie déjà très abondante sur cette méthode d'estimation et les méthodes connexes. Deming et Stephan (1940) comptent parmi les premiers qui ont associé la méthode itérative du quotient à la notion de distance minimum conditionnelle. Les principes fondamentaux de cette approche sont analysés dans Ireland et Kullback (1968). Brackstone et Rao (1979) ont bien décrit l'application de cette méthode à la correction des poids de sondage. La série de méthodes comprend aussi deux fonctions de critère qui se rattachent à la méthode multiminimale du maximum de vraisemblance. La relation entre cette méthode et la méthode itérative du quotient a fait l'objet de nombreuses études voir, par exemple, Bishop, Fienberg et Holland (1976). Fienberg (1986) souligne que les critères de distance analysés dans le présent article peuvent être considérés comme des cas spéciaux d'une famille de fonctions paramétriques analysées dans Cressie et Read (1984).

2. MÉTHODES FONDÉES SUR LA NOTION DE DISTANCE MINIMUM CONDITIONNELLE

2.1 Méthodes fondées sur les poids des ménages

Considérons un échantillon de K ménages, dont les poids initiaux sont définis par le vecteur $\bar{S} = (S_1, \dots, S_K)$. Dans cet article, S_k désigne l'inverse de la probabilité de sélection du k ième ménage; dans certaines applications, le poids initial peut inclure d'autres facteurs de correction, par exemple des facteurs de non-réponse.

Une classe de méthodes utilisant des chiffres de population dans la pondération des ménages

CHARLES H. ALEXANDER¹

RÉSUMÉ

On analyse une série de méthodes fondées sur la notion de "distance minimum conditionnelle" qui assurent la concordance des poids des ménages avec les données supplémentaires concernant l'âge, le sexe, la race et le statut marital. Les poids conditionnels se rapprochent autant que possible des poids initiaux, qui sont fondés sur l'inverse de la probabilité de sélection. Cette série de méthodes comprend la méthode itérative du quotient et la méthode des moindres carrés généralisés même que la méthode multivariée du maximum de vraisemblance, où les cellules de la distribution présentent des catégories de ménages. À l'aide de modèles d'observation simples, nous étudions les propriétés de ces méthodes dans une situation de sous-dénombrement systématique des catégories de ménages. Après une comparaison avec la méthode de la personne principale, nous concluons à la nécessité de mieux connaître la nature du sous-dénombrement avant de décider de la méthode la plus appropriée. Les méthodes fondées sur la distance minimum conditionnelle ou méthode de la personne principale).

NOTES CLÉS: Pondération; information supplémentaire; méthode itérative du quotient; méthode de la personne principale; champ d'enquête.

1. INTRODUCTION

La post-stratification est souvent utilisée pour rajuster les poids de sondage en fonction des données supplémentaires sur le nombre d'unités de certaines catégories dans la population. Par exemple, on peut obtenir des estimations indépendantes de la population pour diverses cellules de post-stratification portant sur l'âge, l'origine raciale et le sexe en rajustant les chiffres du recensement en fonction des variations démographiques observées depuis le recensement. Ces estimations indépendantes sont souvent appelées "chiffres de population". Avant la stratification à posteriori, chaque personne (ou ménage) de l'échantillon a déjà un poids qui correspond habituellement à l'inverse de sa probabilité de sélection. Dans la post-stratification, chaque poids est multiplié par un facteur de correction sous forme de ratio qui s'applique à chaque cellule de sorte que la somme des poids corrigés pour une cellule soit égale à son chiffre de population. Cette correction est particulièrement importante lorsqu'il y a un sous-dénombrement systématique des ménages ou des personnes à l'intérieur des ménages.

Dans la plupart des enquêtes démographiques du U.S. Census Bureau, la post-stratification sert à attribuer des poids aux personnes de l'échantillon mais non aux ménages, car il est plus difficile d'obtenir des estimations indépendantes pour les ménages. Pour attribuer des poids aux ménages dans ces enquêtes, on utilise plutôt une version de la méthode de la personne principale. Selon la version fondamentale de cette méthode, on attribue aux ménages des poids équivalents à celui attribué par post-stratification à la personne principale du ménage. Dans la section 2, nous exposons les règles qui permettent de définir la personne principale d'un ménage. En utilisant les poids attribués aux personnes après la post-stratification, la méthode de la personne principale permet d'incorporer les estimations indépendantes du nombre de personnes dans les poids attribués aux ménages.

La pondération ont exigé un temps d'utilisation du processeur environ trois fois plus long dans le cas de la méthode MCG que dans le cas de la MIQ. La méthode MCG a nécessité aussi plus de temps pour la mise en mémoire des fichiers. (Les matrices utilisées pour la CPS ont des dimensions très appréciables, par exemple \bar{P} , P_0 , X , et \bar{N} comptent environ 14,000 lignes pour chaque groupe de renouvellement.)

5. RÉSUMÉ ET CONCLUSIONS

Cet article avait pour but de comparer la MIQ et la méthode MCG dans le cadre de la CPS tant au point de vue des macro-données qu'au point de vue des micro-données. En ce qui concerne les macro-données, les deux méthodes produisent des estimations comparables. Les mesures de proximité ont permis de constater que les corrections apportées aux poids d'échantillon pour respecter les contraintes de contrôle étaient légèrement inférieures dans le cas de la MIQ. En revanche, celle-ci avait tendance à produire des mesures de proximité légèrement plus élevées (par rapport aux MCG) pour les sous-agrégats de minorités ethniques. Là où les deux méthodes diffèrent le plus l'une de l'autre a trait aux corrections qui ont apportées aux groupes de population qui sont surdéterminés ou sousdéterminés. Cette analyse nous permet de déduire que la MIQ exige moins de temps machine que la méthode MCG pour la correction du second degré de la CPS.

REMERCIEMENTS

Les auteurs tiennent à exprimer leur reconnaissance à Fritz Scheuren, qui a révisé la version originale de cet article, de même qu'aux arbitres et au rédacteur associé qui, par leurs commentaires très utiles, ont contribué à améliorer la qualité de cet article.

BIBLIOGRAPHIE

- BAILLAR, B. (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23-30.
- DEMING, W.E., et STEPHAN, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11, 427-444.
- HANSON, R.H. (1978). The Current Population Survey design and methodology. Technical Paper 40, U.S. Bureau of the Census.
- HUANG, E.T., et FULLER, W.A. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of the Section on Social Statistics, American Statistical Association*, 300-305.
- IRELAND, C.T., et KULLBACK, S. (1968). Contingency tables with given marginals. *Biometrika*, 55, 179-188.
- LUERY, D. (1986). Weighting sample survey data under linear constraints on the weights. *Proceedings of the Section on Social Statistics, American Statistical Association*, 325-350.
- NEYMAN, J. (1949). Contribution to the Theory of the X^2 Test. Dans *Proceedings of the First Berkeley Symposium on Mathematical Statistics and Probability*, (éd. J. Neyman), Berkeley: University of California Press, 239-273.
- WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.
- ZIESCHANG, K.D. (1986). A generalized least squares weighting system for the Consumer Expenditure Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 64-71.

Tableau 3

Comparaison des corrections apportées par la MIQ et les MCG, 1984

Proportion des enregistrements dans l'intervalle pour lesquels MIQ/MCG <0.95 ou >1.05	Proportion de l'échantillon pour laquelle MIQ/MCG <0.95 ou >1.05	Proportion de l'échantillon total	Intervalle de taux de couverture	Catégorie de contrôle
0.219	0.057	0.007	<0.7	Âge/sexe/
0.136	0.116	0.022	0.7-0.8	origine
0.019	0.147	0.241	0.8-0.9	raciale
0.019	0.504	0.699	0.9-1.1	
0.084	0.069	0.021	1.1-1.2	
0.275	0.106	0.010	>1.2	Âge/sexe/
0.198	0.078	0.010	<0.7	origine
0.058	0.032	0.014	0.7-0.8	ethnique
0.033	0.135	0.106	0.8-0.9	
0.022	0.741	0.869	0.9-1.1	
0.202	0.007	0.001	1.1-1.2	
0.373	0.007	0.001	>1.2	Etat
0.031	0.068	0.056	<0.7	
0.042	0.180	0.111	0.7-0.8	
0.030	0.325	0.278	0.8-0.9	
0.018	0.342	0.479	0.9-1.1	
0.009	0.009	0.026	1.1-1.2	
0.040	0.077	0.049	<1.2	

Ce rapport décrit la relation entre les corrections apportées au poids d'une personne de l'échantillon par la MIQ d'une part et par la méthode MCG d'autre part. Pour les besoins de la comparaison, nous dirons qu'un rapport inférieur à 0.95 ou supérieur à 1.05 représente un écart significatif entre les corrections apportées par l'une et l'autre des deux méthodes. Pour chaque série de chiffres de population, nous avons calculé des ratios E/C (taux de couverture), où E est l'estimation d'échantillon fondée sur les poids des personnes de l'échantillon avant la stratification a posteriori et C est le chiffre de population. Pour chaque catégorie de contrôle (état, âge/sexe/origine ethnique, âge/sexe/origine raciale), nous avons classé les enregistrements de l'échantillon selon le taux de couverture. Le tableau 3 donne la répartition de l'échantillon selon l'intervalle de taux de couverture et la valeur du rapport MIQ/MCG , de même que la proportion des enregistrements dans chaque intervalle de taux de couverture pour lesquels le rapport MIQ/MCG est inférieur à 0.95 ou supérieur à 1.05. Les données de ce tableau indiquent que pour chaque catégorie de contrôle, la MIQ et la méthode MCG étaient plus susceptibles de produire des corrections divergentes dans le cas d'enregistrements tirés de groupes de population qui avaient été surdénumbrés ou sousdénumbrés jusqu'à un certain point dans l'enquête (c'est-à-dire pour lesquels le taux de couverture n'était pas près de 1) que dans le cas d'enregistrements tirés de groupes de population qui avaient été dénumbrés normalement.

4.3 Ressources informatiques

L'application des deux méthodes a été exécutée sur un Système 370 d'IBM au National Institute of Health au moyen de PROC MATRIX du système SAS. La préparation des fichiers

Tableau 2

Comparaison des mesures de proximité fondée sur 8 groupes de renouvellement pour chaque année (nombre de GR pour lesquels $MIQ < MCG$)

	M_A		M_B	
	1983	1984	1983	1984
Total	8	8	4	7
De race blanche	7	7	3	4
De race noire	3	3	1	1
D'origine hispanique	0	0	0	0
Hommes	2	7	1	5
Femmes	8	8	8	8

Bien que M_B devrait normalement être minimisée par la méthode des MCG, la valeur de M_B fondée sur les poids calculés selon cette méthode pour l'échantillon total était plus élevée que la valeur de M_B fondée sur les poids calculés selon la MIO dans 11 cas sur 16.

En cherchant à expliquer cette contradiction apparente, nous avons remarqué qu'après six itérations, les estimations produites par la MIO ne concordent toujours pas avec les chiffres de population pour l'âge, le sexe et l'origine ethnique. Le degré de "non-convergence" est toutefois *très faible* (moins de 1.0% pour toutes les catégories de contrôle). Cependant, compte tenu de l'écart entre les valeurs de M_B fondées respectivement sur la MIO et les MCG, il suffirait de modifier les poids calculés selon la MIO dans une proportion de 0.1 à 0.2% pour obtenir le résultat inverse. En soumettant la MIO à 15 itérations, qui n'ont pas suffi néanmoins à assurer la convergence, nous avons pu constater que les résultats observés pour M_B pouvaient être attribuables à la légère non-convergence de la MIO. (Il convient de souligner que la méthode des MCG minimise M_B pour la série de méthodes de correction qui produisent des estimations conformes aux chiffres de population. Comme les estimations produites par la MIO lors de la CPS ne concordent pas avec les chiffres de population, elle ne fait pas partie de la catégorie précitée.)

Bien qu'une méthode de correction comme la MIO ou les MCG puisse minimiser une mesure de proximité pour l'échantillon global, il n'est pas dit qu'elle fera la même chose pour des sous-agrégats de l'échantillon qui ont été comparés à des chiffres de population (par exemple, personnes de race noire, personnes d'origine hispanique, personnes de sexe masculin). Compte tenu de l'utilisation de chiffres de population et du fait que la mesure globale de proximité est minimisée, il semblerait souhaitable d'avoir une méthode de correction qui produise aussi de faibles mesures de proximité pour les sous-agrégats. La méthode des MCG a produit de telles mesures dans presque tous les groupes de renouvellement pour les personnes de race noire et dans plusieurs groupes de renouvellement pour les personnes de

Comparaison des corrections

La MIO et la méthode MCG définissent toutes deux des facteurs de correction pour les cellules formées par l'intersection des contraintes marginales. Le même facteur s'applique à tous les enregistrements d'une cellule. Afin de comparer les corrections apportées aux poids par l'une ou l'autre des deux méthodes, nous avons comparé les facteurs définis par chacune des méthodes pour chaque enregistrement de l'échantillon à l'aide du rapport suivant:

$$RRE/GLS = [(W_{2i}/W_{1i})^{RRE}] / [(W_{2i}/W_{1i})^{GLS}].$$

Comme l'indiquent les données du tableau 1, aucune différence ou tendance notable ne se dégage des estimations pondérées de la population active ou des estimations de l'erreur type calculées par l'une et l'autre des deux méthodes lorsque ces estimations sont sous-agrégées en fonction du sexe et de l'origine raciale ou ethnique.

En ce qui concerne les estimations de la population active selon la combinaison sexe/origine raciale ou ethnique, les différences relatives estimées (en valeur absolue) étaient toutes inférieures à 0,3% (ce qui est bien au-dessous des CV estimés de chaque estimation). Pour la majorité de ces estimations, notamment celles ayant trait à l'ensemble des personnes de sexe masculin ou de sexe féminin ou aux personnes de race blanche, la différence relative en valeur absolue était inférieure à 0,1%.

Pour un bon nombre de caractéristiques, le signe de la différence relative a changé de 1983 à 1984; l'écart entre les estimations produites par les deux méthodes ne semble donc pas suivre une tendance particulière.

En ce qui concerne les estimations de l'erreur type pour les estimations nationales de la population active, les différences relatives en valeur absolue étaient toutes inférieures à 1,9% pour la population totale, à 0,7% pour les personnes de race blanche, à 3,5% pour les personnes de race noire et à 3,1% pour les personnes d'origine hispanique.

b. Indices de l'effet du nombre de mois d'inclusion dans l'échantillon

De nombreuses études montrent que les estimations calculées à l'aide des poids finals de la CPS sont entachées d'un biais relatif qui varie selon le nombre de mois d'inclusion du groupe de renouvellement dans l'échantillon (Bailar 1975). Des indices de l'effet du nombre de mois d'inclusion dans l'échantillon

$$I_k = (8Y_k/Y) \times 100,$$

ont été calculées pour juillet 1983 et juillet 1984 d'après les estimations calculées par la MIQ et les MCG.

Dans les deux cas, on a obtenu des indices virtuellement identiques pour la population active selon l'origine raciale, selon le sexe et selon l'origine ethnique.

4.2 Micro-données

a. Correction des poids d'échantillon

La MIQ aussi bien que les MCG minimisent une mesure de proximité des poids d'échantillon (poids initial vs poids corrigé). Dans le cas de la MIQ, la mesure est (Ireland et Kullback 1968)

$$M_A = \sum_i W_{2i} \ln (W_{2i}/W_{1i}).$$

Pour les MCG, la mesure est (Luery 1986)

$$M_B = \sum_i (W_{2i} - W_{1i})^2 / W_{1i}$$

où W_{1i} = poids de l'enregistrement i de l'échantillon avant correction, W_{2i} = poids de l'enregistrement i de l'échantillon après correction.

La comparaison des mesures de proximité (voir le tableau 2) donne des résultats intéressants et parfois étonnants. Ainsi, la MIQ a produit des valeurs moindres pour les deux mesures. La méthode des MCG avait tendance à produire des valeurs moindres pour certains sous-groupes, notamment les personnes de race noire et les personnes d'origine hispanique. Il convient de souligner que l'écart entre les valeurs produites par les deux méthodes était presque toujours inférieur à 1%.

Estimations de la population active selon la combinaison sexe/origine raciale ou ethnique

[illegible]

où \bar{F} = vecteur ($n \times 1$) des poids finals calculés (W_z) pour chacune des n personnes de l'échantillon,

\bar{P} = vecteur ($n \times 1$) des poids des personnes de l'échantillon avant la stratification a posteriori (W_{1i}),

\bar{P}_0 = matrice diagonale ($n \times n$) dont la diagonale est formée des W_{1i} ,

X = matrice descriptive ($n \times k$), où les lignes correspondent aux personnes de l'échantillon et les colonnes aux cellules de contrôle. La matrice est formée de zéros (0) et de uns (1), qui indiquent les catégories de contrôle appropriées pour chacune des n personnes de l'échantillon,

\bar{N} = vecteur ($k \times 1$) des estimations démographiques supplémentaires, qui correspond aux colonnes de X . Ces estimations servent aussi à la MIQ dans la CPS.

Les colonnes de X doivent être linéairement indépendantes pour qu'il existe un inverse de la matrice ($X' P_0 X$). Lors de la formation des matrice X et \bar{N} pour la CPS, les 137 cellules de contrôle utilisées pour la MIQ (état, âge/sexe/origine ethnique, âge/sexe/origine raciale) ont été réduites à $k = 132$ cellules qui étaient linéairement indépendantes.

La seule solution de $X' \bar{F} = \bar{N}$ qui minimise $f(\bar{F})$ est (voir Luery 1986)

$$\bar{F} = \bar{P} + P_0 X (X' P_0 X)^{-1} (\bar{N} - X' \bar{P})$$

Bien qu'il ne soit pas nécessaire que les éléments de \bar{F} soient positifs, ils l'étaient tous dans ce cas sans que l'on ait eu à poser des conditions additionnelles. Huang et Fuller (1978) et Zieschang (1986), entre autres, analysent des méthodes qui permettent d'obtenir des poids non négatifs dans ce contexte.

4. RÉSULTATS

4.1 Macro-données

a. Estimations

Les estimations de la population active ont été établies pour plusieurs groupes démographiques pour juillet 1983 et juillet 1984 au moyen des poids finals calculés par la MIQ et les MCG. On a aussi calculé les erreurs types pour les deux méthodes au moyen d'un estimateur de la méthode des groupes aléatoires défini comme suit (Wolter 1985):

$$\sum_{k=1}^8 (8Y_k - \bar{Y})^2 / 56,$$

où Y_k = somme des poids des enregistrements de l'échantillon pour le k -ième groupe de renouvellement avec la caractéristique Y ,
 \bar{Y} = somme des Y_k .

On a utilisé cet estimateur de variance, même s'il ne tient pas compte du plan de sondage à plusieurs degrés de la CPS, parce que le fichier de micro-données à grande diffusion de la CPS ne contenait aucune information sur le plan de sondage.
On a calculé des différences relatives pour les estimations de niveau et les estimations d'erreur type. La formule de la différence relative était

$$(Y_{GLS} - Y_{RRE}) / Y_{RRE},$$

où Y_{MIQ} = estimation de Y fondée sur les poids calculés selon la MIQ,

Y_{MCG} = estimation de Y fondée sur les poids calculés selon la méthode des MCG.

2) Correction par quotient selon la combinaison âge/sex/origine ethnique:

$$n_{ijk}^{(1,2)} = (m_{j.}/n_{j.}) n_{ijk}^{(1,1)} = b_j^{(1)} n_{ijk}^{(1,1)}$$
$$= a_i^{(1)} b_j^{(1)} n_{ijk}^{(1,1)}$$

3) Correction par quotient selon la combinaison âge/sex/origine raciale:

$$n_{ijk}^{(1,3)} = (m_{..k}/n_{..k}) n_{ijk}^{(1,2)} = d_k^{(1)} n_{ijk}^{(1,2)}$$
$$= a_i^{(1)} b_j^{(1)} d_k^{(1)} n_{ijk}^{(1,1)}$$

où $n_{i..}$ = total de ligne pour l'échantillon
 $n_{.j.}$ = total de colonne pour l'échantillon
 $n_{..k}$ = total de couche pour l'échantillon.

L'exécution des trois étapes ci-dessus correspond à une itération de la méthode itérative du quotient. On reprend les trois étapes en substituant à chaque fois la valeur de $n_{ijk}^{(h,3)}$ (donnée-échantillon corrigée issue de la troisième étape de la h -ième itération) à $n_{ijk}^{(h,2)}$ dans l'étape 1) jusqu'à ce qu'on ait effectué 6 itérations. (Le nombre d'itérations effectuées dans la CPS a été établi en fonction des propriétés de convergence de la MIQ pour cette enquête et des gains relatifs réalisés selon le nombre d'itérations.) La valeur finale $\{n_{ijk}^{(6,3)}\}$ est définie comme $\{n_{ijk}^{(h,3)}\}$.
Le facteur de correction des poids d'échantillon pour les enregistrements de la cellule $\{ijk\}$ est

$$F_{ijk} = n_{ijk}^{(6,3)} / n_{ijk}^{(h,3)}$$
$$= \prod_{h=1}^6 a_i^{(h)} b_j^{(h)} d_k^{(h)}.$$

Pour obtenir les poids corrigés, on multiplie les poids d'échantillon observés avant l'application de la MIQ par le facteur F_{ijk} approprié.

3. APPLICATION DES MCG DANS LA CPS

La méthode des moindres carrés généralisés (MCG) permet de corriger les poids d'échantillon découlant des étapes de pondération antérieures en minimisant les carrés pondérés des corrections, sous réserve que les poids corrigés satisfassent à une série de contraintes "de contrôle" linéaires. C'est exactement ce à quoi Deming et Stephan cherchaient à s'attaquer lorsqu'ils ont mis au point la MIQ. Comme celle-ci, la méthode MCG produit des MEAN dans certaines conditions, en l'occurrence lorsque toutes les corrections apportées aux poids d'échantillon par l'intermédiaire d'une mesure de proximité (voir sous-section 4.2).
En ce qui a trait à la CPS, chaque dimension qui définit une série de chiffres de population dans la stratification a posteriori courante définita aussi une série de contraintes linéaires pour la méthode MCG. La fonction à minimiser est

$$f(\bar{F}) = (\bar{F} - \bar{P})' P_0^{-1} (\bar{F} - \bar{P})$$
$$= \sum_i (W_{2i} - W_{1i})^2 / W_{1i}$$

à la condition que $X' \bar{F} = \bar{N}$,

et juillet 1984). Afin d'apprécier les différences qui pouvaient exister entre les deux méthodes pour cette application, nous nous sommes intéressés aussi bien aux macro-données qu'aux micro-données.

2. MÉTHODE D'ESTIMATION PAR STRATIFICATION À POSTERIORI UTILISÉE ACTUELLEMENT DANS LA CPS

Dans la CPS, on utilise actuellement la méthode itérative du quotient (MIQ) pour corriger les poids d'échantillon à l'intérieur d'un groupe de renouvellement de manière à ajuster les estimations démographiques de l'échantillon à des estimations démographiques supplémentaires dans trois catégories distinctes (état, âge/sexe/origine ethnique, âge/sexe/origine raciale).

Deming et Stephan (1940) ont été les premiers à proposer l'application de la MIQ pour la correction des données de tableaux de remplacement de la méthode des moindres carrés. On a montré que la MIQ produisait de meilleurs estimateurs asymptotiquement normaux (MEAN) dans un échantillonage aléatoire simple et qu'elle minimisait, comme nous le verrons dans la sous-section 4.2, les corrections apportées aux poids d'échantillon (Irland et Kuillback 1968). De plus, même si elle produit des estimations biaisées, la MIQ peut parfois réduire l'erreur quadratique moyenne des estimations d'enquête. C'est le rôle que l'on attribue à la MIQ dans la CPS (Hanson 1978).

En ce qui a trait à la CPS, la MIQ vise à corriger les chiffres d'échantillon $\{n_{ijk}\}$ décollant des étapes de pondération antérieures afin de produire des chiffres d'échantillon ajustés $\{n_{ijk}^*\}$ à la condition que

$$(A) \quad \sum_{j,k} n_{ijk}^* = m_{i..}$$

$$(B) \quad \sum_{i,k} n_{ijk}^* = m_{.j.}$$

$$(C) \quad \sum_{i,j} n_{ijk}^* = m_{..k}$$

soient satisfaites simultanément,

où i = état ($i = 1, \dots, 51$),

j = âge/sexe/origine ethnique ($j = 1, \dots, 16$),

k = âge/sexe/origine raciale ($k = 1, \dots, 70$),

$m_{i..}$ = estimation supplémentaire relative à l'état,

$m_{.j.}$ = estimation supplémentaire relative à la combinaison âge/sexe/origine ethnique,

$m_{..k}$ = estimation supplémentaire relative à la combinaison âge/sexe/origine raciale.

La MIQ permet de corriger proportionnellement, en se servant de quotients, les données d'échantillon selon les trois catégories de critères (c'est-à-dire, état, âge/sexe/origine ethnique et âge/sexe/origine raciale) par étapes successives.

(1) Correction par quotient selon l'état:

$$n_{ijk}^{(1,1)} = (m_{i..}/n_{i..}) n_{ijk} = a_i^{(1)} n_{ijk}$$

Méthode alternative pour ajuster les estimations de la Current Population Survey aux chiffres de population

K.R. COPELAND, F.K. PEITZMEIER, et C.E. HOY¹

RÉSUMÉ

La Current Population Survey utilise la méthode itérative du quotient dans l'estimation par stratification à posteriori pour redresser les estimations démographiques d'échantillon en fonction d'estimations démographiques fondées sur le recensement. Dans cet article, les auteurs proposent une deuxième méthode, fondée sur les moindres carrés généralisés, et la comparent à la méthode actuelle.

NOTES CLÉS: Moindres carrés généralisés; stratification à posteriori; estimation itérative du quotient.

1. INTRODUCTION

La Current Population Survey (CPS) produit des estimations de la population active pour l'ensemble de la population civile hors institutions d'âge actif aux États-Unis à partir d'un échantillon aléatoire mensuel à plusieurs degrés de quelque 60,000 logements. Chaque mois, un échantillon est interviewé afin de recueillir des données démographiques et des données sur l'activité pour tous les occupants adultes civils des logements échantillonnés.

Les estimations, sous-agrégées selon des caractéristiques démographiques, sont publiées mensuellement. Des estimations concernant d'autres sous-agrégats de la population (états, familles, anciens combattants, salariés, inactifs, etc.) sont aussi publiées mensuellement, trimestriellement ou annuellement.

On calcule le poids des personnes de l'échantillon en utilisant la probabilité de sélection, un facteur de compensation de la non-réponse et un facteur de correction sous forme de quotient afin de réduire la composante de la variance attribuable au prélèvement d'unités primaires d'échantillonnage. On corrige ensuite les poids ainsi obtenus au moyen d'une méthode d'estimation par stratification à posteriori de manière à rapprocher les estimations démographiques de l'enquête d'estimations démographiques supplémentaires. Enfin, les poids ainsi corrigés sont utilisés dans une méthode d'estimation composite, puis désaisonnalisés afin d'obtenir des estimations nationales (Hanson 1978).

Pour certains sous-domaines de population (familles, salariés, inactifs, revenus de la famille et anciens combattants), le calcul d'estimations détaillées requiert des poids d'échantillon déterminés par des méthodes de correction qui viennent s'ajouter à l'estimation par stratification à posteriori.

La méthode des moindres carrés généralisés (MCG) pourrait peut-être remplacer l'estimation par stratification à posteriori ou servir à intégrer les diverses méthodes de correction de la CPS. On a déjà proposé l'application des MCG dans l'enquête sur les dépenses des consommateurs et des analyses ont été faites à ce sujet (Zieschang 1986).

Nous nous proposons dans cet article de comparer la méthode d'estimation par stratification à posteriori utilisée actuellement dans la CPS (méthode itérative du quotient) et la méthode MCG en nous fondant sur les données de la CPS pour deux mois particuliers (juillet 1983

¹ K.R. Copeland, F.K. Peitzmeier et C.E. Hoy, Division of Statistical Methods, Office of Employment and Unemployment Statistics, Bureau of Labor Statistics, Washington, D.C. 20212, États-Unis.

es biographies en démographie a fait l'objet de nombreuses recherches méthodologiques (Courgeau 1984; Haeringer 1972; Riandey 1985). Notre méthode se veut uniquement un outil simple et fiable d'aide à la collecte des données. À chaque utilisateur de retenir les variables classées dans le temps sur la fiche «AGEVEN», et une fois la trame biographique collectée approfondir le(s) domaine(s) étudié(s) à l'aide du questionnaire.

REMERCIEMENTS

Les auteurs tiennent à remercier les arbitres pour leurs commentaires utiles.

BIBLIOGRAPHIE

- NTOINE, Ph., et DIOUF, P.D. (1986). Changements démographiques en milieu urbain. Communication présentée au Séminaire sur la mortalité au Sénégal. Dakar.
- ONNET, D. (1984). Occultation, omissions. Quelques problèmes soulevés par l'enquête quantitative en matière de santé. *Medicus Mundi*, 11.
- OURGÉAU, D. (1984). Relations entre cycle de vie et migrations. *Population*, 39, 483-513.
- OURGÉAU, D., et LELIEVRE, E. (1986). Nuptialité et agriculture. *Population*, 41, 303-326.
- OX, R., et OAKES, D. (1984). *Analysis of Survival Data*. London: Chapman & Hall.
- DIRECTION DE LA STATISTIQUE, 1981: *Enquête Sénégalaise sur la Fécondité, 1978 - Rapport National d'Analyse*, 1.
- ARGUES, Ph. (1985). L'évaluation du niveau de la mortalité à partir des données des enquêtes EMIL. *Les enquêtes sur la mortalité infantile et juvénile (EMIJ)*, 1, 60-84.
- ERRY, B. (1977). Le fichier événement. Une nouvelle méthode d'observation rétrospective. Dans *l'Observation démographique dans les pays à statistiques déficientes*. Liège, Belgique: Ordina Editions, 137-150.
- AERINGER, Ph. (1972). Méthodes de recherche sur les migrations africaines. Un modèle d'interaction biographique et sa transcription synoptique. *Cahiers ORSTOM*, 9, 439-453.
- COTT, Ch. (1985). Les problèmes de déperdition dans les enquêtes suivies. Dans *Les enquêtes sur la mortalité infantile et juvénile (EMIJ)*, 1, 44-47.
- IANDEY, B. (1985). L'enquête "biographie familiale professionnelle et migratoire" (INED, 1981). Le bilan de la collecte. Dans *Migrations internes, collecte des données et méthode d'analyse*. Université de Louvain, 117-134.

Tableau 1

Mortalité selon le lieu de naissance (en pour mille)

	Pikine	Dakar	Autres Villes	Rural	Total	Pkn-Rural Test
Infantile	52	57	45	114	58	-6,586**
Jeune ville	55	62	90	156	68	-10,093**
Effectif	5155	1513	644	704	8016	

Suivant le lieu de naissance de l'enfant, des niveaux différents de mortalité ont été relevés. De nombreuses mères observées à l'enquête, sont des migrantes provenant d'une autre ville ou d'un village de l'intérieur du pays. Parmi leurs enfants, ceux nés en milieu rural présentent des risques de mortalité bien plus grands que ceux nés dans l'agglomération dakaroise. Le quotient de mortalité juvénile (entre 1 et 4 ans) traduit bien les risques liés aux disparités socio-économiques. Le risque de mourir entre 1 et 4 ans est 2,84 fois plus élevé pour les enfants nés dans les villages que pour ceux nés à Pikine. Le "test Z" montre que la différence entre les deux quotients (quotient de mortalité Pikine et quotient de mortalité rural) est significative. On test l'hypothèse que le quotient de mortalité des enfants nés à Pikine est le même que celui des enfants nés en milieu rural. Les tailles d'échantillon étant assez élevées, l'approximation par une loi normale est justifiée. Le test basé sur la statistique Z admet la loi de probabilité d'une normale centrée réduite sous l'hypothèse d'égalité des quotients de mortalité. Le signe "***" indique une différence significative au niveau de $\alpha = 0,05$. Un recueil rétrospectif classique sans distinction du lieu de naissance de l'enfant nous aurait amené à assimiler les naissances hors Pikine à celles de l'agglomération observée et induit un niveau plus élevé de mortalité (mortalité juvénile de 68 pour mille au lieu de 55 pour mille). Par ailleurs, une seconde analyse est possible pour chacune des femmes observées; on peut constituer un fichier biographique simplifié, dont les étapes successives sont délimitées par les naissances. Une relation est ainsi établie entre les événements matrimoniaux, les changements de résidence, et les données génésiques. On peut également reconstruire les principales étapes du cheminement migratoire depuis la naissance du premier enfant, ou bien depuis le mariage. Les données longitudinales recueillies de cette façon se prêtent bien aux méthodes récentes d'analyse des interférences entre phénomènes (Courgeau et Lelievre (1986), Cox et Oakes (1984)).

5. CONCLUSION

L'information recueillie pour chacune des variables est très succincte mais elle doit permettre de dégager quelques différences importantes, et de recueillir des informations sur les conditions de vie à la naissance et au moment du décès. La méthodologie de collecte retenue est adaptée à la collecte des données concernant la vie génésique des femmes et le devenir de leurs enfants. L'intérêt majeur de la fiche « AGBEVEN » réside dans la facilité apportée au repérage dans le temps de certains événements et dans le classement des événements les uns par rapport aux autres, en gardant la possibilité d'insérer les événements omis au fur et à mesure de la conversation. La souplesse d'utilisation de la fiche « AGBEVEN » nous amène à suggérer sa transposition à d'autres domaines comme celui des biographies professionnelles, ou des itinéraires migratoires en mettant en parallèle localité de résidence, profession, situation matrimoniale, situation familiale, conditions de logement... Le domaine de l'analyse

À l'usage cette fiche est incomplète car il manque une question permettant d'éviter les ambiguïtés de confusion entre mort-né et enfant décédé juste après la naissance, confusion entre fausses couches et avortement qu'entre mort-né et décédé juste après la naissance. Certains termes ou mots français ne sont pas directement traduisibles en wolof. C'est ainsi que pour mort-né, il n'y a pas une question qui permette à elle seule d'obtenir la réponse souhaitée. Il faut donc au moins deux questions pour avoir l'information désirée. Pour ce qui est des intervalles intergénéraliques plus ou moins longs, les enquêteurs posent la question suivante par exemple: "Lou am dikhané té Moussa ak Ali?" ("qu'est-ce qu'il y a eu entre Moussa et Ali?"). À cette interrogation, les femmes comprennent à juste titre mort-nés, avortements, fausses couches, etc. Pour avoir une réponse satisfaisante des précisions sont nécessaires: "Dikhané té Moussa ak Ali, amo fi dom diou dé guinaw bou mou indé bakhané?" ("entre Moussa et Ali, n'auriez-vous pas un enfant décédé après qu'il ait manifesté un signe quelconque de vie?"). La confusion est surtout entretenue par le fait que la distinction entre fausse couche et mort-né n'est pas toujours évidente et que l'enfant n'a de prénom qu'après le semaine. Aussi, selon certaines ethnies, ce n'est qu'à partir de ce moment qu'il est vraiment pris en compte. Une colonne précisant si l'enfant a ou non crié au moment de la naissance aurait donc été la bienvenue.

La fiche «AGEVEN» de l'enquête Pikine a certes donné des informations plus satisfaisantes que le graphique utilisé lors de l'enquête sénégalaise fécondité de par la nature et la quantité des informations recueillies. Par contre elle n'a pas pour autant éliminé l'inconvénient de la tendance à arrondir en années les intervalles intergénéraliques (environ 37% des intervalles) et plus particulièrement les intervalles de deux ans, qui représentent environ 20% des intervalles observés entre les naissances successives. De plus, cette technique se révèle insuffisante pour recenser toutes les issues de grossesses des jeunes filles ayant déjà été enceintes, n'ayant eu aucune naissance vivante. Certains biais, certes classiques en démographie, subsistent donc et cette méthode ne disperse d'une grande vigilance sur le terrain.

1. LA TRANSCRIPTION DE LA FICHE «AGEVEN» SUR LE QUESTIONNAIRE ET SON EXPLOITATION INFORMATIQUE

Le questionnaire concernant la vie génésique des femmes a été conçu de façon à permettre la meilleure transcription possible des informations recueillies sur la fiche «AGEVEN». Dans un premier temps, les caractéristiques propres à chacun des enfants sont notées dans l'ordre chronologique des naissances, ainsi que les dates de décès s'il y a lieu. Dans un second temps, l'enquêteur devait relever la situation matrimoniale au moment de chacun des événements afin de suivre les changements éventuels de conjoint. Ensuite étaient prises en compte l'évolution de la situation socio-économique du père, celle de la mère, ainsi que l'évolution des conditions d'habitat et les différents lieux de résidence. L'ensemble de l'enquête comprenait d'autres questionnaires concernant les caractéristiques du ménage, des individus, et des femmes soumises à l'observation.

Le mode de recueil de l'information permet deux types d'analyse. L'une concerne l'analyse classique de la mortalité par génération et par sous-population (selon le quartier, le type de logement, etc. . .). Mais l'intérêt majeur de cette étude est de permettre l'analyse de la mortalité (et de la fécondité) en tenant compte des comportements migratoires et de l'évolution des conditions socio-économiques des femmes soumises à l'enquête. Grâce à cette méthode, la mortalité n'est plus interprétée en fonction des seules conditions socio-économiques au moment de l'enquête, mais au contraire, elles est rapportée aux conditions réellement vécues au moment de l'événement, et l'on peut de cette façon mieux appréhender les différences propres aux conditions de vie en milieu urbain (Pikine dans notre cas).

me suis mariée il y a 15 ans). Enfin, grâce à cette fiche des événements datés de façon imprécise, peuvent être situés: tel enfant est né entre celui né le 10-2-74 et celui né en 1978. Il est fort probable que l'enfant soit alors né durant l'année 1976. Cette fiche nécessite que l'enquêteur porte un regard critique sur l'enchaînement des événements et qu'il cherche à la compléter au maximum en s'assurant de la fiabilité et de la cohérence des réponses portées. Cela n'est possible qu'en instaurant un dialogue confiant avec la personne enquêtée.

Après avoir enregistré toutes les naissances vivantes déclarées par l'enquêtée, l'enquêteur s'intéresse alors aux intervalles intergénéstiques. Il arrive que certains de ces événements n soient pas rapportés dans les premières réponses, mais l'utilisation de la fiche «AGEVEN» permet à l'enquêteur de mieux dépitster les événements omis. Aussi, interroge-t-il l'enquêtée sur ce qui s'est passé, à chaque fois qu'il s'aperçoit que l'intervalle entre deux naissances vivantes est supérieur à deux ans. Les réponses fournies par l'enquêtée permettent ainsi d relever les avortements, mort-nés, naissances suivies de décès, voire d'obtenir des informations sur les pratiques contraceptives. Ce point ne faisait pas partie des objectifs de l'enquête mais le dialogue qui se noue autour de l'«AGEVEN» peut cependant permettre d'approfondir des questions ayant trait à la planification des naissances.

Chacun de ces événements étaient mis en relation avec le lieu, la situation matrimoniale et le partenaire de la femme au moment de l'événement. Après avoir relevé tous les événements ayant affecté la femme, l'enquêteur devait estimer de façon beaucoup plus précise la date de naissance de la mère. En effet, au moment de remplir le questionnaire "ménage", l'enquêteur avait déjà inscrit la date de naissance de la mère qui lui avait été communiquée soit par la femme ou par le chef de ménage. Cette fois, seul avec l'enquêtée et en possession des événements ayant affecté la femme, il était à même de fournir le meilleur âge possible de l'enquêtée.

Prenons l'exemple d'une femme. Awa est née en 1956 à Kaolack: elle déclare avoir trois enfants, Ibrahim, qui aurait aujourd'hui 10 ans, né à Dakar, décédé à l'âge de 4 ans à Pikine Abdoul, né le 5 janvier 1978 à Dakar, et Aminata née le 18 décembre 1984 à Pikine. Cette femme s'est mariée une première fois à l'âge de 17 ans à Thiès. Elle a divorcé en 1979 (elle résidait alors à Pikine). Elle s'est remariée en 1982 et résidait à cette époque à Pikine (figure 2). En menant l'interview l'enquêteur constatera un écart de presque 7 ans entre Abdoul et Aminata. Il devra insister pour savoir si durant cet intervalle d'autres naissances ou d'autres grossesses ne se sont pas produites. Dans ce cas précis, le divorce et le remariage trois ans plus tard peuvent expliquer cet écart, mais il faut s'assurer auprès de la femme qu'un tel écart ne masque pas des événements démographiques.

La forme interactive prise par l'interview semble favoriser le dialogue avec la personne enquêtée et améliorer le contact entre enquêteur et enquêtée, malheureusement trop souvent basé sur le doute de l'enquêteur et la méfiance de l'enquêtée (Bonnet 1984). Au fil et à mesure que l'enquêteur ou l'enquêtée poursuit son investigation de nouveaux événements sont déclarés. Lorsque l'on demande s'il n'y a pas un événement entre deux naissances à plus de deux ans d'intervalle, la personne interrogée s'étonne et formule l'un ou l'autre type de réponse: si elle n'a connu aucun événement "Pourquoi demandez-vous cela?". Par contre, si un événement omis existe, bien souvent elle questionne "Qui est-ce qui te l'a dit?". En ayant l'impression que l'enquêteur sait déjà quelque chose. La fiche «AGEVEN» apparaît comme un instrument de "divination" tels les cauris. Parfois, l'entretien revêt un aspect ludique, la personne interrogée est contente de pouvoir remettre de l'ordre dans le déroulement d'événements passés; une femme à la vie gènesique et matrimoniale complexe a même voulu une copie de sa fiche «AGEVEN». Comme toutes les enquêtes l'utilisation de cette fiche se heurte à certaines difficultés. Il est parfois difficile ou délicat de s'isoler avec la personne enquêtée et bien souvent les femmes sont gênées si la fiche met en évidence des événements concernant un partenaire précédant le mari actuel.

BNR/ORSTOM

CODES		AGEVEN		BNR/ORSTOM	
N _i + date + nom + lieu = naissance enfant n° i D _i + date + nom + lieu = décès enfant n° i M _i + date + nom + lieu = mariage n° i D _i + date + nom = divorce conjoint n° i V _i + date + nom = veuvage conjoint n° i FC = fausse couche A = avortement MN = mort-né P _i = père n° i					
IDENTIFIANT		Nom		Ménage:	
		Ilot:		Concession:	
		N° de la femme:			

sur la fécondité. Le nom est le même, les possibilités d'utilisation diffèrent. La fiche «AGEVEN» dont un exemple est présentée à la figure 2 comprend trois colonnes:

- la première concerne les événements démographiques (naissance (N); décès (DCD); avortement (AV); fausse couche (FC); mort-né (MN)). Chaque événement, naissance ou décès doit être suivi de son rang chronologique, des nom et prénom de l'enfant et éventuellement de la date précise;
- la seconde colonne concerne les événements matrimoniaux et le rang de chacun des partenaires ou conjoints (mariages (M), divorces (D), veuvages (V), le rang des différents père indicé P1, P2, ..., Pn).
- la troisième colonne permet de noter le lieu de résidence lors de chacun des événements démographiques et matrimoniaux. Cette colonne permet d'établir ensuite l'itinéraire migratoire des femmes et de déterminer leur date d'installation à Pikine.

La fiche «AGEVEN» est un outil méthodologique qui remplit plusieurs fonctions:

- repérer les événements dans le temps;
- aider la femme à situer temporellement des événements dont elle a oublié la date;
- s'assurer de l'exhaustivité des événements démographiques vécus par la femme interrogée;
- repérer les changements de résidence et de localiser les événements;
- vérifier la cohérence des événements entre eux.

Le déroulement de l'entretien comprenait deux phases, l'une concernant le ménage, l'autre les femmes de 15-49 ans. Le questionnaire "ménage", au niveau duquel sont répertoriées toutes les personnes résidentes ou non du ménage, s'intéresse en particulier à la filiation de personnes recensées, à leur lien de parenté avec le chef de ménage ou de noyau, à leur sexe à leur situation matrimoniale, à leur date de naissance ou leur âge. Le questionnaire "femmes de 15-49 ans" concerne l'ensemble des femmes, résidentes et présentes dans le ménage, âgées de 15-49 ans. La fiche «AGEVEN» est utilisée au moment de remplir ce questionnaire.

Pour transcrire les informations obtenues sur cette fiche l'enquêteur peut prendre différents points de référence (la date de naissance de la femme; la date de la première naissance...), et reconstitue avec la participation de l'enquêtée toute sa ligne de vie, c'est-à-dire toutes les autres événements qui ont affecté sa vie tels mariage, divorce, et les différents accouchements. Cette opération se décompose comme suit:

1. Après enregistrement de la première naissance vivante, d'enquêteur demande à l'enquêtée de lui communiquer toutes les naissances vivantes qui ont suivi dans un ordre croissant, que l'enfant soit décédé ou non, qu'il vive dans le ménage ou ailleurs.
2. Ensuite l'enquêteur enregistre ces naissances sur la fiche en se servant des pièces officielles qui lui sont présentées. Les pièces officielles existaient dans la plupart des cas pour les enfants surtout s'ils étaient nés dans la région de Dakar. Pour l'âge des mères, par contre, comme pour certaines autres naissances d'enfants, l'enquêteur devrait se servir d'éléments provenant du calendrier historique pour déterminer les dates (mois et années).

La fiche «AGEVEN» permet de situer les événements soit selon l'âge de la femme au moment de l'événement, ou la durée écoulée depuis l'événement, soit selon la date de cet événement. Tout écart important entre deux naissances, ou toute incohérence entre deux événements est ainsi repéré plus facilement au cours de l'entretien avec la femme.

A l'aide de la fiche «AGEVEN», on peut également vérifier la cohérence des événements. Ainsi, deux enfants ne peuvent naître à moins de neuf mois d'intervalle; ou bien une femme ne peut déclarer s'être mariée à l'âge de 12 ans, avoir eu son premier enfant à 14 ans et en 1970 et déclarer être né en 1950. Dans ce cas, il est fort probable que la date de naissance de la femme soit erronée et il faudra la corriger.

La fiche permet d'enregistrer aussi bien des événements donnés avec une date précise, qu'un événement dont on peut simplement donner l'âge (tel enfant a aujourd'hui 10 ans; j'ai

LV = Ligne de vie de la femme
DE = Durée écoulée depuis l'événement

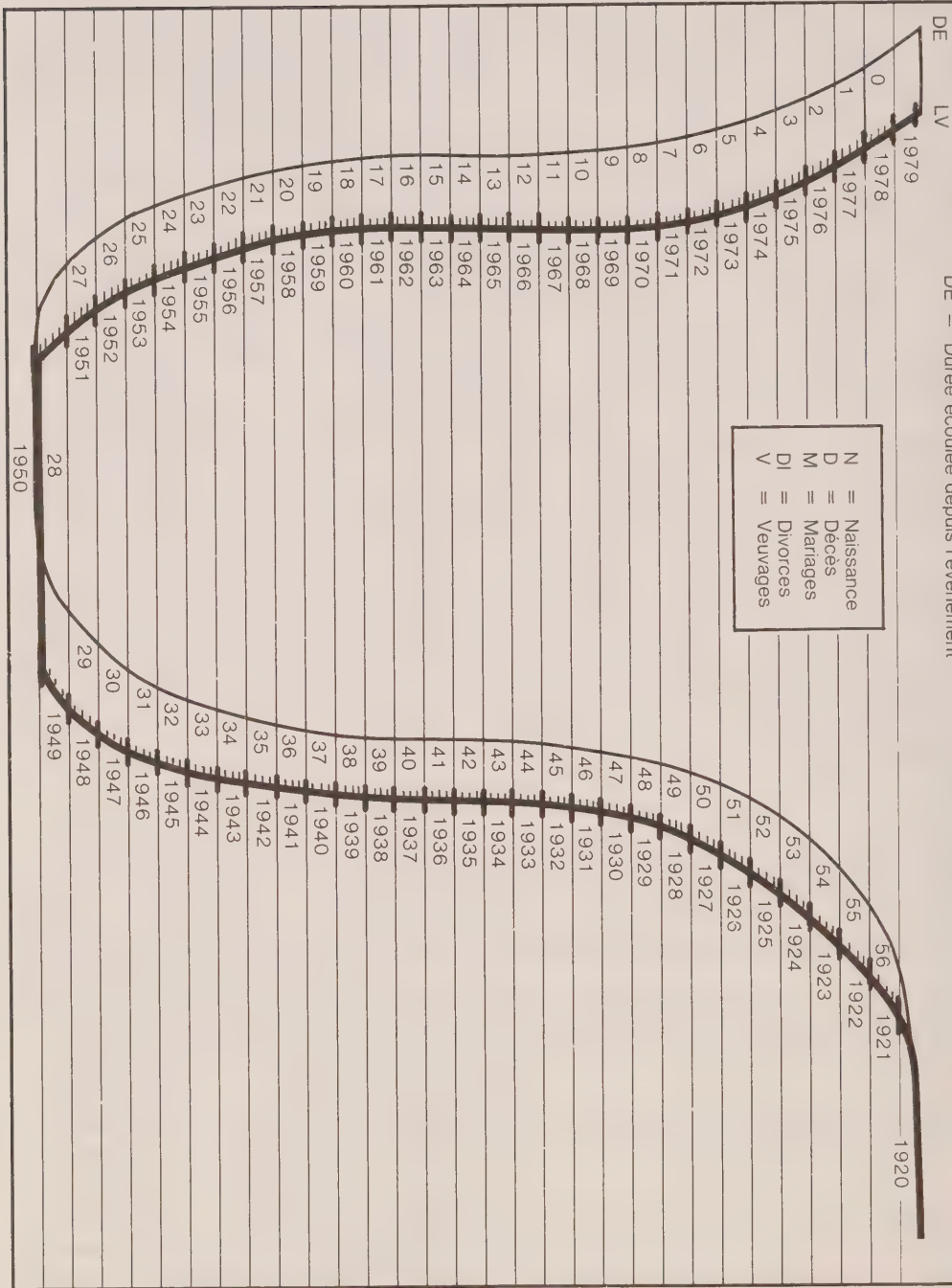


Figure 1. Graphique AGEVEN de l'enquête Sénégalaise sur la fécondité

- urbaine et les changements de comportements démographiques. Ces objectifs devaient guider notre stratégie de collecte;
- obtenir une liste complète des événements observés (principalement naissances et décès) - dater le plus précisément possible ces événements
- replacer les événements dans leur contexte socio-économique du moment (situation matrimoniale, situation professionnelle du mari et de la femme, conditions de vie).

2. RECUEIL ET DATATION DES ÉVÉNEMENTS DÉMOGRAPHIQUES

Réussir une enquête rétrospective nécessite en particulier d'établir une biographie la plus précise possible (concernant le domaine étudié) de chaque personne soumise à l'enquête; il faut donc trouver le moyen d'aider la personne interrogée à replacer dans le temps les événements vécus dans le passé.

Plusieurs améliorations méthodologiques ont déjà été proposées. Ainsi Ferry (1977) a utilisé un "fichier événement" qui repose sur l'affection d'une fiche à chaque événement. Selon l'auteur "toute l'originalité de la méthode consiste, avec la personne enquêtée elle-même, à mettre de l'ordre dans la succession des événements (Grossesses, unions et ruptures d'union, lieux d'habitation, etc. . .) et à les situer les uns par rapport aux autres (. . .). La technique consiste alors à retrouver avec la personne enquêtée la succession, la logique, les interférences et finalement la biographie individuelle". Toutefois, cette méthode est relativement complexe et nécessite le maniement de nombreuses fiches sur le terrain et lors de l'exploitation. Un autre mode de classement et de datation des événements a été retenu lors de l'enquête sénégalaise sur la fécondité de 1978, il s'agit du graphique «AGEVEN» (âge à l'événement): L'utilisation du graphique «AGEVEN» pour l'enquête sénégalaise visait deux buts:

- permettre une meilleure estimation de l'âge des femmes et celui de leurs enfants grâce à une datation relativement précise;
 - permettre une bonne estimation de la fécondité en faisant l'historique des maternités de toutes les femmes.
- Le graphique «AGEVEN» de l'enquête sénégalaise fécondité (figure 1) se présente sous la forme de deux courbes. La courbe de droite figurant la ligne de vie de la femme (courbe L.V.) est représentée par un intervalle gradué en trimestres qui, permettait de situer dans l'année les événements affectant la femme. La courbe de gauche appelée D.E. donnait la durée écoulée entre le moment de l'événement et la date de l'enquête. Aussi, à chaque année sur la courbe L.V. correspondait un âge sur la courbe D.E., et inversement. Ce graphique, repris par l'enquête ivoirienne sur la fécondité, apparaît surtout comme un instrument de datation des événements.

3. L'UTILISATION DE LA FICHE «AGEVEN» LORS DE L'ENQUÊTE MENÉE À PIKINE

Nous avons cherché à combiner certains des avantages de chacune de ces méthodes de collecte: d'une part la simplicité d'emploi du graphique «AGEVEN» dans la datation des événements, et d'autre part la possibilité de saisir plusieurs types d'événements et leur classement les uns par rapport aux autres comme le permet le fichier événement. Nous avons donc systématisé cette fiche «AGEVEN» en distinguant le repérage des événements démographiques (naissances, décès), les changements de situation matrimoniale, et les changements de lieu de résidence. Par commodité nous avons retenu le même vocable pour désigner notre fiche, que celui l'employé pour dénommer le graphique utilisé lors de l'enquête sénégalaise

La fiche «AGEVEN»: un outil pour la collecte des données rétrospectives

PHILIPPE ANTOINE, XAVIER BRY et PAP DEMBA DIOUF¹

RÉSUMÉ

La fiche «AGEVEN» permet, grâce à sa simplicité d'emploi, une meilleure datation des événements et d'opérer un classement respectif des événements démographiques (naissances et décès), des changements matrimoniaux et des changements de lieux de résidence. Les données obtenues servent à reconstruire avec précision les conditions socio-économiques au moment où se produisent les événements démographiques étudiés.

MOTS CLÉS: Enquête rétrospective; biographies; enquête démographique.

1. INTRODUCTION

Deux grandes méthodes de collecte sont à la disposition du démographe pour recueillir des données afférentes au mouvement naturel (natalité et mortalité): l'observation suivie et le questionnaire rétrospectif. La méthode d'observation suivie (suivre un même échantillon de population pendant un intervalle de temps relativement long) est en théorie celle qui donne les résultats les plus précis. Elle présente toutefois certains inconvénients. Les coûts d'enquête sont élevés du fait des nombreux passages nécessaires à l'observation. Les délais d'obtention des résultats demeurent relativement longs. Enfin, en milieu urbain, l'application de la méthode se heurte à l'extrême mobilité de la population ce qui entraîne une déperdition importante de l'échantillon, comme celle rencontrée lors des enquêtes mortalité infantile et juvénile (EMIJ) de l'IFORD (Scott 1985, Fargues 1985).

La méthode rétrospective donne des résultats moins fiables car elle fait davantage appel à la mémoire des enquêtes. Cependant elle porte sur une période totale d'observation en général plus longue que celle des enquêtes longitudinales mises en place ces dernières années dans les pays africains. Les risques d'omission des événements demeurent élevés et leur datation reste imprécise; enfin, en milieu urbain, le recueil de la vie passée mêle des événements qui se sont déroulés dans la ville objet de l'enquête et d'autres, plus anciens, qui se sont produits dans d'autres lieux de résidence (urbain ou rural).

Désirant connaître la mortalité et la fécondité différentielles à Pikine, dans la banlieue de Dakar, et souhaitant obtenir rapidement des résultats assez fiables, notre choix s'est porté sur une méthode de collecte permettant de reconstituer avec précision les facteurs de risques de mortalité infantile au moment du décès pour chacun des enfants des femmes soumises à l'enquête. L'enquête a été réalisée conjointement par la Direction de la Statistique du Sénégal et l'Orstom (Antoine et Diouf 1986). L'enquête de terrain s'est déroulée de mars à mai 1986. Les premiers résultats étaient disponibles dès septembre 1986. Cette méthode distingue donc de la méthode rétrospective la plus employée qui ne prend en compte que les caractéristiques socio-économiques et culturelles des femmes au moment de l'enquête, alors que celles-ci ont pu considérablement évoluer au cours de leur vie féconde (amélioration ou dégradation des conditions d'habitat, changement dans la situation matrimoniale, dans l'activité. . .). Cette méthode permet de mieux évaluer les interférences entre l'insertion

¹ Philippe Antoine, démographe et Xavier Bry, statisticien à l'ORSTOM BP 1386 Dakar Sénégal; Pap Demba Diouf, démographe à la direction de la Statistique BP 116 Dakar Sénégal.

- KEYFITZ, N. (1957). Estimates of sampling variance where two units are selected from each stratum. *Journal of the American Statistical Association*, 52, 503-510.
- PLATEK, R., et SINGH, M.P. (1976). Méthodologie de l'enquête sur la population active du Canada. No. 71-526 au catalogue, Statistique Canada.
- RAO, J.N.K. (1975). Unbiased variance estimation for multi-stage designs. *Sankhyā*, Série C, 37, 133-139.
- RAO, J.N.K., HARTLEY, H.O., et COCHRAN, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society*, 24, 482-490.
- SINGH, M.P., et DREW, J.D. (1981). Redesigning continuous surveys in a changing environment. *Techniques d'enquête*, 7, 44-73.
- WOODRUFF, R.S. (1971). A simple method for approximating the variance of a complicated estimator. *Journal of the American Statistical Association*, 66, 411-414.

Tableau 6B

Taux de couverture des intervalles de confiance de 95% établis à l'aide de l'estimateur de la variance pour une seule itération

Caractéristique	Province (N.-E.)					
	RE	RE	RE	RE	RMR	Halifax
Personnes occupées	94.5	92.8	94.0	94.7	94.7	94.9
Personnes en chômage	92.1	90.7	91.4	91.8	92.7	92.7
Personnes actives	96.2	93.0	93.6	95.2	95.2	96.0

Nous avons aussi constaté que l'estimation par le quotient calculée en deux itérations est presque non biaisée, le biais maximum dans les 21 cas étant de 0.35%.

4. CONCLUSIONS

En ce qui a trait aux caractéristiques de la population active considérées dans cette étude, nous avons vu que la méthode de Keyfitz appliquée à des estimations sans correction par le quotient (dans ce cas la méthode de Keyfitz est réduite à une méthode répétitive) produit des estimations qui ont un biais positif très élevé et qui sont peu efficaces tandis que les autres méthodes appliquées au même genre d'estimations produisent des estimations qui sont plus efficaces et dont le biais est négligeable.

Toutefois, lorsqu'appliquées à des estimations par le quotient, toutes les méthodes sans exception produisent des estimations qui ont un biais négligeable. Nous avons aussi montré que l'on pouvait accroître sensiblement l'efficacité de la méthode de Keyfitz au point de la rendre comparable à celle des autres méthodes en augmentant le nombre d'échantillons répétés. Nous avons vu à l'aide de données authentiques de l'EPA que le fait d'utiliser 6 groupes de renouvellement au lieu de 2 pseudo-échantillons répétés dans la formule de la variance de Keyfitz n'introduisait pas de biais de renouvellement. Comme en font foi les résultats de l'étude de Monte Carlo, l'estimateur de la variance pour une seule itération, déduit de la méthode de Keyfitz par une linéarisation de Taylor, produit des estimations assez justes de la variance des estimations par le quotient calculées en deux itérations et est aussi caractérisé par de bons taux de couverture.

REMERCIEMENTS

Les auteurs tiennent à exprimer leur reconnaissance aux deux arbitres ainsi qu'au rédacteur en chef et à un rédacteur associé pour leurs commentaires utiles sur la version préliminaire de cet article.

BIBLIOGRAPHIE

ARORA, H. R., et BRACKSTONE, G. J. (1977a). An investigation of the properties of raking ratio estimators: I with simple random sampling. *Techniques d'enquête*, 3, 62-83.

ARORA, H. R., et BRACKSTONE, G. J. (1977b). An investigation of the properties of raking ratio estimators: II with cluster sampling. *Techniques d'enquête*, 3, 232-252.

BRACKSTONE, G. J., et RAO, J. N. K. (1979). An investigation of raking ratio estimators. *Sankhyā*, Série C, 41, 97-114.

La formule (3.10) donne la variance des estimations de caractéristiques de la population active pondérées en fonction de M_2 et nécessite l'utilisation des deux poids M_0 et M_2 .

3.3 Application de la formule de la variance établie dans la sous-section 3.2 aux estimations par le quotient calculées en deux itérations

L'application répétée de la méthode de linéarisation de Taylor peut servir à déterminer la formule de la variance pour des estimations par le quotient calculées en deux itérations. La formule obtenue de cette façon est toutefois très complexe. Nous avons supposé que la formule de la variance pour des estimations par le quotient calculées en une seule itération produirait une estimation assez juste de la variance des estimations calculées en deux itérations. Cette hypothèse repose sur le fait que les poids ne subissent que de légères variations après la première itération. Or, la formule de la variance pour les estimations calculées en une seule itération utilise la paire de poids (W_0, W_2) . Nous avons donc décidé de substituer (W_0, W_4) à (W_0, W_2) après avoir constaté que l'utilisation de W_4 ne changeait rien aux coefficients de variation (CV) des estimations des caractéristiques de la population active fondées sur W_4 . La formule de la variance qui utilise la paire de poids (W_0, W_4) sera désignée comme l'estimateur de la variance pour une seule itération.

1. Estimation par le quotient calculée en deux itérations $Y^{(4)}$:

1. Estimation par le quotient calculée en deux itérations $Y^{(4)}$.
2. Estimation de la variance $V(Y^{(4)})$, à l'aide de l'estimateur de la variance pour une seule itération, et estimation du CV correspondant.
3. Intervalle de confiance de 95% (c'est-à-dire, $Y^{(4)} \pm 1.96 \sqrt{V(Y^{(4)})}$).

A la fin de la simulation, nous avons fait la moyenne des 1 000 estimations de CV et avons comparé cette moyenne au CV de Monte Carlo, qui est très près de la valeur réelle. Les résultats sont reproduits dans le tableau 6A. L'écart est inférieur à 8% dans les 21 cas (3 caractéristiques pour chacune des 7 régions) et inférieur à 4% dans 13 cas.

Nous avons aussi déterminé la proportion des intervalles de confiance qui renferment la valeur réelle de la caractéristique. Les résultats sont reproduits dans le tableau 6B. Les taux de couverture pour les "personnes occupées" et les "personnes actives" sont, en règle générale, très près du seuil théorique tandis que ceux pour les "personnes en chômage" sont quelque peu inférieurs mais encore acceptables.

Tableau 6A
CV moyens obtenus par l'estimateur de la variance pour une seule itération et CV de Monte Carlo

Caractéristique	RE 210	RE 220	RE 230	RE 240	RE 250	RMR Halifax	Province (N.-E.)
Personnes occupées	3,52	3,46	3,14	3,05	1,96	2,01	1,08
Personnes en chômage	10,36	12,28	13,13	13,43	10,35	10,55	5,27
Personnes actives	2,98	3,17	2,85	2,73	1,77	1,83	0,91
CV de Monte Carlo							
Personnes occupées	3,48	3,35	2,95	2,86	1,97	1,99	1,11
Personnes en chômage	10,90	12,71	13,28	13,37	11,12	11,31	5,59
Personnes actives	2,76	3,08	2,76	2,53	1,72	1,74	0,92

un secteur spécial peut être incluse dans plus d'une région infrarégionale et la somme-
ion ($\Sigma^{s\geq h}$) sert à additionner toutes les valeurs ($D^{sh}_{(0)}$) comprises dans cette strate. Posons

Alors, (3.7) devient

$$V(Y_{(2)}) = V \left(\sum_h \sum_{n_h} D^{sh}_{(0)} \right) \quad (3.8)$$

Les variables ($\Sigma_l D^{hl}_{(0)}$) sont indépendantes puisqu'elles reposent sur des sous-poids. En
tenant pas compte de la CPF (correction pour population finie), nous pouvons estimer
à variance à l'aide de la formule suivante:

$$V(Y_{(2)}) = \sum_h \frac{n_h}{n_h - 1} \sum_{l=1}^h (D^{hl}_{(0)} - \bar{D}^{hl}_{(0)})^2 \quad (3.9)$$

Or, cette expression implique des espérances mathématiques qui sont inconnues. Nous pouvons
obtenir une approximation assez juste de la variance en remplaçant les espérances mathémati-
ques par leurs estimations; ainsi, de l'équation (3.9) nous déduisons la forme finale de V :

$$V = \sum_h \sum_{n_h} \frac{n_h}{n_h - 1} (D^{hl}_{(2)} - \bar{D}^{hl}_{(2)})^2 \quad (3.10)$$

$$D^{hl}_{(2)} = \sum_{s \geq h} D^{shl}_{(2)}$$

$$\bar{D}^{hl}_{(2)} = \frac{1}{n_h} \sum_{n_h} n_h$$

$$D^{shl}_{(2)} = \sum \left(Z^{shl}_{(2)} - R^{shl}_{(2)} Z^{shl}_{(2)} \right)$$

$$Z^{shl}_{(2)} = \frac{P^{shl}_{(2)}}{P^{shl}_{(0)}} \left(P^{shl}_{(0)} \frac{D^{shl}_{(0)}}{P^{shl}_{(0)}} - Y^{shl}_{(0)} \frac{D^{shl}_{(0)}}{P^{shl}_{(0)}} \right)$$

$$Y^{shl}_{(2)} = Y^{shl}_{(2)} \frac{D^{shl}_{(0)}}{P^{shl}_{(0)}} - Y^{shl}_{(2)} \frac{D^{shl}_{(0)}}{P^{shl}_{(0)}}$$

$$R^{shl}_{(2)} = \frac{Y^{shl}_{(2)}}{Y^{shl}_{(0)}} \frac{D^{shl}_{(0)}}{P^{shl}_{(0)}} = \frac{Y^{shl}_{(2)}}{Y^{shl}_{(0)}} \frac{D^{shl}_{(0)}}{P^{shl}_{(0)}}$$

où s désigne une RMR ou une RE, ou encore la partie d'une RE qui ne correspond pas à des RMR, et P_s est la population de la région intraprovinciale s . En remplaçant $Y_{(1)}^a$ et $P_{(1)}^a$ dans l'équation (3.2) par les équations (3.3) et en appliquant l'approximation de Taylor du premier ordre aux rapports des estimations pondérées en fonction de W_0 , nous obtenons:

$$V(Y_{(2)}) \approx V \left[\sum_a \frac{P_a}{P} E(P_{(1)}^a) \sum_s \frac{E(P_{(0)}^s)}{P_s} \left\{ \left(Y_{(0)}^{sa} - R_{(0)}^{Ysa} P_{(0)}^s \right) \right\} \right] \quad (3.4)$$

$$R_{(0)}^{Ysa} = \frac{E(Y_{(0)}^{sa})}{E(P_{(0)}^{sa})} \text{ et } R_{(0)}^{Psa} = \frac{E(P_{(0)}^s)}{E(P_{(0)}^{sa})}.$$

où

L'équation (3.4) peut être réécrite en fonction d'estimations pour échantillons répétés.

Définissons

$$Z_{(0)}^{Ysha} = \frac{P_a}{P} \frac{E(P_{(1)}^a)}{P_s} \frac{E(P_{(0)}^s)}{P_s} (Y_{(0)}^{sha} - R_{(0)}^{Ysa} P_{(0)}^s),$$

$$Z_{(0)}^{Psha} = \frac{P_a}{P} \frac{E(P_{(1)}^a)}{P_s} \frac{E(P_{(0)}^s)}{P_s} (P_{(0)}^{sha} - R_{(0)}^{Psa} P_{(0)}^s), \quad (3.5)$$

où h désigne une strate de s et i désigne un échantillon répété dans h .

Alors, nous pouvons réécrire (3.4) en modifiant l'ordre des sommations:

$$V(Y_{(2)}) \approx V \left\{ \sum_s \sum_{hes} \sum_{n_h} \sum_a \left(Z_{(0)}^{Ysha} - R_{(1)}^{Ya} Z_{(0)}^{Psha} \right) \right\}$$

$$= V \left(\sum_h \sum_{hes} \sum_{n_h} \sum_{shi} D_{(0)}^{shi} \right) \quad (3.6)$$

où

$$D_{(0)}^{shi} = \sum_a \left(Z_{(0)}^{Ysha} - R_{(1)}^{Ya} Z_{(0)}^{Psha} \right).$$

Pour les strates de secteurs réguliers, les variables $(\sum_{n_h}^i D_{(0)}^{shi})$ sont indépendantes parce qu'elles reposent sur des sous-poids. Toutefois, en ce qui concerne les strates de secteurs spéciaux, ces variables sont fortement corrélées parce que les mêmes enregistrements sont attribués à toutes les régions intraprovinciales désignées.

Nous pouvons réécrire (3.6) comme suit:

$$V(Y_{(2)}) \approx V \left(\sum_{hes} \sum_h \sum_{n_h} \sum_{shi} D_{(0)}^{shi} \right) = V \left(\sum_h \sum_{l=1}^h \sum_{s \geq h} D_{(0)}^{shi} \right) \quad (3.7)$$

où $\sum_{s \geq h}$ désigne la sommation pour toutes les régions intraprovinciales qui renferment la strate h . S'il s'agit d'une strate d'un secteur régulier, la sommation $(\sum_{s \geq h})$ est superflue puis-que la strate ne se trouve que dans une seule région intraprovinciale. En revanche, la strate

itératives par le quotient pour l'échantillonnage aléatoire simple d'unités ou de grappes. Nous utilisons cette méthode pour l'échantillonnage stratifié à plusieurs degrés avec PPT en nous fondant sur l'approche de Woodruff.

Soient $Y_{(0)}$, $Y_{(1)}$, $Y_{(2)}$ les estimations d'une caractéristique y de la population active dans une province, ces estimations étant fondées respectivement sur W_0 , W_1 , et W_2 . Les chiffres entre parenthèses correspondent aux indices inférieurs de W .

Nous pouvons alors exprimer $Y_{(2)}$ de la façon suivante:

$$Y_{(2)} = \sum_a \frac{Y_{(1)}^a}{P_{(1)}^a} P_a \quad (3.1)$$

où $Y_{(1)}^a$ = estimation de la caractéristique y pondérée avec W_1 pour le groupe d'âge-sexe a dans la province;

$P_{(1)}^a$ = estimation de la population pondérée avec W_1 pour le groupe d'âge-sexe a dans la province;

P_a = estimation auxiliaire (ou externe) de la population pour le groupe d'âge-sexe a dans la province.

Posons $F_a = Y_{(1)}^a / P_{(1)}^a$. L'approximation de Taylor du premier ordre pour F_a à $(E(Y_{(1)}), E(P_{(1)}^a))$ est

$$F_a \approx \frac{E(Y_{(1)}^a)}{E(P_{(1)}^a)} + \frac{E(P_{(1)}^a)}{1} \left\{ Y_{(1)}^a - E(Y_{(1)}^a) \right\} - \frac{E(Y_{(1)}^a)}{E(P_{(1)}^a)} \left\{ P_{(1)}^a - E(P_{(1)}^a) \right\}$$

où E désigne l'espérance mathématique.

Nous pouvons alors formuler une approximation de Taylor pour la variance de $Y_{(2)}$:

$$V(Y_{(2)}) = V \left(\sum_a F_a P_a \right) \approx V \left\{ \sum_a \frac{E(P_{(1)}^a)}{P_a} (Y_{(1)}^a - R_{(1)}^a P_{(1)}^a) \right\} \quad (3.2)$$

$$R_{(1)}^a = \frac{E(Y_{(1)}^a)}{E(P_{(1)}^a)}.$$

Or, il est possible d'exprimer les estimations $Y_{(1)}^a$ et $P_{(1)}^a$, fondées sur W_1 , en fonction des estimations pondérées selon W_0 :

$$Y_{(1)}^a = \sum_s \frac{Y_{(0)}^a}{Y_{(0)}^{sa}} P_s,$$

$$P_{(1)}^a = \sum_s \frac{P_{(0)}^a}{P_{(0)}^{sa}} P_s, \quad (3.3)$$

Tableau 5

Comparaison des estimations de la variance pour les secteurs AR par suite de l'utilisation de 2 et de 6 échantillons répétés par strate selon des données de l'EPA pour des RMR mars 1985 - février 1987

Caractéristique	Rapport moyen des moyennes des variances		Rapport moyen des E.T. des variances (2 échantillons/6 échantillons)
	Personnes occupées	Personnes en chômage	Personnes actives
	0.997	0.995	1.003
	1.813	1.515	1.833

Note: Pour chaque RMR, on a calculé les moyennes et les écarts types des estimations de la variance pour 2 et pour 6 échantillons répétés, à partir de données s'étendant sur une période de 24 mois. On a ensuite calculé, pour chaque RMR, le rapport (2 échantillons/6 échantillons) des moyennes des variances et le rapport des E.T. des variances. Les rapports moyens qui figurent dans le tableau sont la moyenne des rapports des 24 RMR.

Les itérations se traduisent en une suite de corrections: premièrement, on ajuste le sous-poids en fonction de la population intraprovinciale (c'est-à-dire, population de RMR ou de régions autres qu'une RMR dans les RE), puis on multiplie le poids ainsi obtenu par le facteur de correction âge-sexe pour la province (le remaniement de l'échantillon a fait passer le nombre de groupes d'âge-sexe de 38 à 24). On répète l'opération une fois pour obtenir une deuxième paire de poids. Il convient de souligner que lorsqu'on définit des cellules de correction pour les RE, on ne tient pas compte des RMR contenues dans ces régions de sorte que les cellules de correction au niveau intraprovincial s'excluent mutuellement. Désignons par W_0 le sous-poids et par (W_1, W_2) et (W_3, W_4) les deux paires de poids tirées respectivement de la première et de la seconde itération. On utilise W_4 pour estimer les caractéristiques de la population active. À cause de l'ordre des corrections, la somme des valeurs de W_4 pour les groupes d'âge-sexe de niveau provincial est égale aux estimations auxiliaires des groupes correspondants mais la somme des mêmes valeurs au niveau intraprovincial (RE et RMR) n'est pas tout à fait égale aux estimations auxiliaires correspondantes. Les écarts sont toutefois très faibles.

Les bases de secteurs spéciaux, qui sont composées des établissements militaires, des institutions et des régions éloignées, ne respectent pas en général les divisions des RE et des RMR et, de ce fait, ne sont pas traitées de la même façon dans l'itération que les bases de secteurs réguliers. Chaque secteur spécial constitue une strate dans chaque province, les seules exceptions étant les régions éloignées du Québec et de l'Alberta, qui sont sous-stratifiées. Les RE et les RMR qui sont utilisées pour la base de secteurs spéciaux sont dites "désignées". Les enregistrements d'un secteur spécial contenus dans le fichier de l'échantillon sont reproduits pour chaque RE ou RMR désignée avec des sous-poids ajustés en fonction de la proportion de la population du secteur contenue dans la RE ou la RMR en question. On applique ensuite la méthode itérative de la façon définie précédemment.

3.2 Formule de la variance pour des estimations par le quotient calculées en une seule itération

Dans cette sous-section, nous déterminons la formule de la variance pour des estimations par le quotient calculées en une seule itération. Pour cela, nous procédons essentiellement à une application répétée de l'approximation de la série de Taylor aux estimations itératives par le quotient jusqu'à ce que nous obtenions une forme linéaire de sous-poids. Nous appliquons ensuite la formule d'itération à la manière de Woodruff (1971). Arora et Brackstone (1977a,b) et Rao (1979) ont aussi recouru à l'application répétée de l'approximation de la série de Taylor pour déterminer la formule de la variance des estimations

Tableau 4

Taux de couverture des intervalles de confiance de 95% pour les estimations de totaux des caractéristiques de la population active avec correction par le quotient

Taux de couverture			
Caractéristique	$P^1_{(R)}$	$P^2_{(R)}$	$P^3_{(R)}$
Personnes occupées	93.6	95.4	94.6
Personnes en chômage	94.3	95.1	95.3
Personnes actives	93.2	95.3	94.6
			94.2

Nous avons analysé cet aspect de la question pour les trois caractéristiques de la population active en calculant, à l'aide de la formule élaborée dans la section 3, les estimations de la variance pour 2 et pour 6 échantillons répétés fondés sur les données de l'EPA pour la période de 24 mois (mars 1985 - février 1987). Nous avons ensuite calculé, pour chaque an, la moyenne et l'écart type (ET) des 24 estimations pour chaque caractéristique de la population active. Enfin, nous avons établi le rapport des moyennes et des écarts types calculés pour chaque plan (2 échantillons/6 échantillons) pour 24 régions métropolitaines de recensement (RMR) et avons fait la moyenne des rapports pour ces 24 régions. Les résultats présentés figurent dans le Tableau 5. Nous pouvons en tirer les conclusions suivantes:

() Le fait d'utiliser 6 échantillons répétés au lieu de 2 a très peu d'effet sur les variances; l'utilisation de groupes de renouvellement influe donc peu sur le biais des estimations de la variance.

() Comme prévu, les variances sont plus stables avec 6 échantillons répétés qu'avec 2 et les résultats ne sont pas très différents de ceux de l'étude de Monte Carlo (voir la première colonne du tableau 3B).

En conclusion, l'utilisation de 6 échantillons répétés accroît sensiblement l'efficacité de méthode de Keyfitz sans influencer de façon notable sur le biais.

3. ESTIMATION DE LA VARIANCE POUR DES ESTIMATIONS ITÉRATIVES PAR LE QUOTIENT

1 Estimations itératives par le quotient dans l'EPA

L'ancien plan de sondage de l'EPA utilisait l'estimation par le quotient pour domaines post-stratifiés. Le sous-poids, qui est le résultat de la correction du poids de base en fonction de la non-réponse, était rajusté par le quotient en fonction d'estimations auxiliaires (ou ex-trnes) de la population cible de l'EPA pour 38 post-strates selon l'âge et le sexe au niveau provincial. La population cible de l'EPA comprend toutes les personnes âgées de 15 ans et us, sauf les membres des forces armées, les pensionnaires d'institution et les personnes vivant ans les réserves indiennes.

L'estimation par le quotient avait pour effet de relever sensiblement la qualité des données provinciales alors que les données infraprovinciales demeuraient peu fiables. Afin d'acquies la fiabilité des données infraprovinciales, surtout au niveau des régions économiques (E) et des RMR, on a appliqué une technique d'estimation itérative par le quotient, qui permet de corriger simultanément des estimations aux niveaux provincial et infraprovincial.

Biais (en pourcentage) des estimateurs de la variance des estimations de totaux des caractéristiques de la population active sans correction par le quotient

Tableau 2A

Caractéristique	V_1	V_2	V_3	V_4
Personnes occupées	23.4	24.5	-4.7	-6.3
Personnes en chômage	6.3	6.6	3.7	1.2
Personnes actives	24.2	25.2	-5.1	-6.7

Biais (en pourcentage) des estimateurs de la variance des estimations de totaux des caractéristiques de la population active avec correction par le quotient

Tableau 2B

Caractéristique	$V_1^{(R)}$	$V_2^{(R)}$	$V_3^{(R)}$	$V_4^{(R)}$
Personnes occupées	3.7	4.3	-1.1	-3.1
Personnes en chômage	5.3	5.5	4.0	1.4
Personnes actives	4.5	5.0	-0.5	-2.5

Efficacité de V_2, V_3 et V_4 par rapport à V_1
(Eff. rel. de $V_j = [EQM(V_1)/EQM(V_j)]^{1/2}, j = 2, 3, 4$)

Tableau 3A

Caractéristique	V_2	V_3	V_4
Personnes occupées	1.51	3.22	3.11
Personnes en chômage	1.52	1.71	1.76
Personnes actives	1.49	3.24	3.12

Efficacité relative

Efficacité de $V_2^{(R)}, V_3^{(R)}$ et $V_4^{(R)}$ par rapport à $V_1^{(R)}$
(Eff. rel. de $V_j^{(R)} = [EQM(V_1^{(R)})/EQM(V_j^{(R)})]^{1/2}, j = 2, 3, 4$)

Tableau 3B

Caractéristique	$V_2^{(R)}$	$V_3^{(R)}$	$V_4^{(R)}$
Personnes occupées	2.13	2.59	2.52
Personnes en chômage	1.57	1.71	1.76
Personnes actives	2.08	2.56	2.51

Nous avons aussi évalué ces estimateurs pour des estimations par le quotient basé sur la population de l'ensemble des strates. Par ailleurs, nous avons calculé de tels estimateurs, désignés par $V_{(R)}^j$, $j = 1, 2, 3, 4$, pour chaque échantillon de Monte Carlo en appliquant la méthode de linéarisation de Taylor, puis nous avons déterminé le biais (en pourcentage) de ces quatre estimateurs (Tableau 2B) et l'efficacité des trois derniers par rapport au premier (Tableau 3B).

Nous constatons que les estimateurs 1 et 2 ont un biais beaucoup moins élevé pour des estimations par le quotient, surtout en ce qui a trait aux personnes occupées et aux personnes actives. Les estimateurs 3 et 4 ont aussi un biais moindre en ce qui concerne ces deux caractéristiques mais un biais à peu près inchangé pour ce qui est des personnes en chômage. Rien que les biais des quatre estimateurs soient faibles, seul le biais de l'estimateur 3 pour les personnes actives s'est révélé non significatif à un seuil de 5%. En ce qui concerne les parts observées entre les biais, seuls ceux se rapportant aux estimateurs 1 et 2 se sont avérés non significatifs à un seuil de 5% pour les trois caractéristiques.

Par ailleurs, nous nous sommes servis des quatre estimateurs étudiés pour déterminer les intervalles de confiance (IC) de 95% pour les estimations par le quotient de chaque échantillon de Monte Carlo. Le taux de couverture des intervalles correspondait à la proportion des IC qui renfermaient la valeur réelle du total de la caractéristique. Les résultats figurent dans le Tableau 4 et montrent que les quatre estimateurs produisent des résultats très satisfaisants pour toutes les caractéristiques. Comme les taux de couverture indiqués dans le Tableau 4 se rapprochent sensiblement du seuil de confiance théorique, les faibles biais observés dans le Tableau 2B n'ont plus vraiment d'importance. Nous pouvons en conclure qu'en ce qui concerne le biais, les quatre estimateurs de la variance ne sont pas très différents des autres pour des estimations par le quotient. Par ailleurs, l'efficacité relative des estimateurs 3 et 4 n'est plus que très légèrement supérieure à celle de l'estimateur 2, peu importe la caractéristique. En l'occurrence, ces trois estimateurs ont une efficacité relative supérieure à 2 pour les personnes occupées et les personnes actives. Pour ce qui a trait aux personnes en chômage, l'efficacité relative est quelque peu inférieure dans les trois cas, les estimateurs se situant entre 1,5 et 1,8, ce qui est très comparable aux rapports observés pour la même caractéristique dans le cas des estimations non corrigées. Il convient de souligner ici que l'estimateur 1 est calculé avec 19 degrés de liberté (soit 1 par strate). Par contre, les trois autres estimateurs sont calculés avec 81 degrés de liberté puisque chaque UPE est un échantillon répété. Nous pouvons en conclure qu'une augmentation du nombre d'échantillons répétés aura pour effet d'accroître sensiblement la stabilité de l'estimateur de la variance et Keyfitz pour les estimations par le quotient et de la rendre comparable à celle des deux autres estimateurs (voir Tableau 3B).

4. Estimateur de la variance de Keyfitz avec 2 échantillons répétés par rapport à 6 échantillons répétés pour l'EPA

Les résultats de l'étude de Monte Carlo ont montré que la méthode de Keyfitz se compare avantageusement, au point de vue du biais et de l'efficacité, aux autres méthodes d'estimation de la variance pour des estimations par le quotient lorsque chaque méthode utilise le même nombre d'échantillons répétés. En outre, la méthode de Keyfitz a l'avantage d'être simple tandis qu'estimer la variance de variations ou de moyennes à l'aide des autres méthodes est un exercice très complexe. On a donc décidé d'étendre la méthode de Keyfitz aux secteurs R. Dans le but d'accroître l'efficacité de cette méthode, on a remplacé les deux échantillons répétés de l'ancien plan de sondage par 6 groupes de renouvellement qui tiennent en compte d'échantillons répétés. Au départ, on craignait principalement que cette substitution se traduise par une forte hausse de l'estimation de la variance imputable au biais de renouvellement.

Maintenant, définissons

$$Y = \sum_{h=1}^{19} Y_h, \\ Y_t = \sum_{h=1}^{19} Y_{ht}, \\ Y_{jt} = \sum_{h=1}^{19} Y_{jht}, \quad j = 1, 2, 3, 4,$$

où $t = 1, 2, \dots, 1000$.

Y_t est l'estimation du total Y établie à partir du t -ième échantillon de Monte Carlo et Y_{jt} , $j = 1, 2, 3, 4$ désigne tour à tour les quatre estimateurs de la variance correspondants. L'espérance mathématique et la variance de Monte Carlo, désignées respectivement par E^* et V^* , sont définies ainsi pour T échantillons de Monte Carlo:

$$E^*(\theta) = \frac{1}{T} \sum_{t=1}^T \theta_t, \\ V^*(\theta) = \frac{1}{T} \sum_{t=1}^T [\theta_t - E^*(\theta)]^2,$$

où θ est un estimateur du paramètre inconnu θ et θ_t est l'estimation tirée du t -ième échantillon. De ces définitions nous déduisons la variance de Monte Carlo de l'estimateur X , $V^*(X)$ de même que l'espérance mathématique et la variance de Monte Carlo de l'estimateur d la variance V_j , soit $E^*(V_j)$ et $V^*(V_j)$ respectivement, pour $j = 1, 2, 3, 4$.

Définissons maintenant le biais de l'estimateur de la variance V_j par l'expression suivante

$$B_j = E^*(V_j) - V^*(V_j),$$

et le biais en pourcentage par:

$$PB_j = 100 \frac{B_j}{V^*(V_j)}, \quad j = 1, 2, 3, 4.$$

Alors, l'erreur quadratique moyenne (EQM) de V_j est définie par:

$$MSE_j = V^*(V_j) + B_j^2, \quad j = 1, 2, 3, 4.$$

Nous définissons l'efficacité de V_j par rapport à l'estimateur de la variance de Keyfitz avec deux échantillons répétés (c'est-à-dire, V_1) par l'expression suivante:

$$\text{Eff. Rel.}(V_j \text{ vs. } V_1) = (EQM_1/EQM_j)^{1/2}, \quad j = 2, 3, 4.$$

Dans cette étude, nous considérons trois caractéristiques de la population active: personnes occupées, personnes en chômage et personnes actives. Les tableaux 2A et 3A donnent respectivement les biais et les efficacités relatives des estimateurs de la variance pour les trois caractéristiques. En ce qui concerne le biais, nous constatons que l'estimateur 1 ressemble à l'estimateur 2 tandis que l'estimateur 3 ressemble à l'estimateur 4. Les estimateurs 1 et 2 ont des biais positifs très élevés, notamment en ce qui concerne les personnes occupées et les personnes actives, tandis que les estimateurs 3 et 4 ont des biais relativement faibles. Au point de vue de l'efficacité, les estimateurs 3 et 4 s'équivalent et sont très supérieurs aux estimateurs 1 et 2. De plus, l'estimateur 2 est plus efficace que l'estimateur 1.

Tableau 1
Strates utilisées pour l'étude de Monte Carlo

Strate	Nombre de logements	Nombre d'UPE	Nombre d'UPE sélectionnés	Taille d'échantillon espérée
1	737	49	6	29.5
2	490	33	4	19.6
3	745	45	6	29.8
4	720	34	6	28.8
5	621	37	6	24.8
6	630	38	6	25.2
7	503	31	4	20.1
8	340	23	4	13.6
9	472	33	4	18.9
10	468	33	4	18.7
11	367	28	4	14.7
12	390	23	4	15.6
13	626	36	6	25.0
14	650	39	6	26.0
15	350	22	4	14.0
16	736	46	6	29.4
17	573	35	6	22.9
18	773	48	6	30.9
19	866	64	8	34.6
Total	11,057	697	100	442.3

2.3 Etude de Monte Carlo

Afin d'évaluer le biais des quatre estimateurs définis ci-dessus et leur stabilité relative, nous avons effectué une étude de Monte Carlo avec 19 strates de l'enquête sur la population active de la région métropolitaine de recensement (RMR) de Halifax en nous servant de données du recensement de 1981. Pour les besoins de cette étude, nous nous sommes servis des données de l'échantillon de recensement auquel était destiné le questionnaire complet, en l'occurrence un échantillon systématique de 20% sélectionné dans les secteurs de dénombrement. Nous avons fixé le taux de sondage, $1/W$, à 0.04 pour que la taille espérée de l'échantillon soit la même que dans la version remaniée de l'EPA. Nous avons déterminé un nombre pair de groupes aléatoires dans chaque strate de manière que la taille espérée de l'échantillon dans ces groupes soit le plus près possible de 4.5 pour assurer la conformité avec le plan actuel de l'EPA. Le tableau 1 donne les 19 strates choisies pour l'étude ainsi que le nombre d'UPE, le nombre d'UPE sélectionnés, le nombre de logements, la taille espérée de l'échantillon et les totaux correspondants pour les 19 strates. Mille échantillons ont été produits individuellement dans chacune des 19 strates par la méthode de Monte Carlo; on s'est servi à cette fin du plan de sondage décrit dans la sous-section 2.1 (plan fondé sur la méthode des groupes aléatoires).

Soit Y_{hi} l'estimation du total Y_h pour la strate h , tirée du i -ième échantillon de Monte Carlo, $h=1, 2, \dots, 19$, et $i=1, 2, \dots, 1,000$. De même, Y_{jht} , $j=1, 2, 3, 4$ désigne tour à tour les quatre estimateurs de la variance de Y_{hi} .

(2) Estimateur de la variance de Rao, Hartley et Cochran (1962)

La formule de cet estimateur repose sur l'hypothèse que le nombre d'unités secondaires m_i qui doivent être sélectionnées suivant un échantillonnage aléatoire simple (EAS). L'estimateur de la variance est défini par

$$V_3(Y) = A \sum_n^I \pi_i \left(\frac{M_i}{m_i} \frac{1}{Y_i} - \frac{1}{Y} \right)^2 + \sum_n^I \frac{\pi_i}{d_i} M_i^2 \left(\frac{1}{m_i} - \frac{1}{Y} \right) s_i^2 \quad (2.8)$$

ou

$$A = \frac{\sum_n^I N_i^2 - N^2}{N^2 - \sum_n^I N_i^2}, \quad (2.9)$$

$$s_i^2 = \frac{1}{m_i - 1} \sum_{k=1}^k (y_{ik} - \bar{y}_i)^2. \quad (2.10)$$

M_i est le nombre de logements contenus dans l'UPÉ sélectionnée dans le i -ième groupe parmi lesquels m_i logements sont sélectionnés suivant un échantillonnage systématique. Toutefois, l'estimation de la variance est calculée suivant l'hypothèse d'un EAS. La valeur de y pour le k -ième logement tiré de l'UPÉ qui a elle-même été sélectionnée dans le i -ième groupe est y_{ik} et $\bar{y}_i = y_i / m_i$.

Comme $\pi_i / d_i = W_i / W_i$ et $M_i / m_i = W_i$ (ces égalités ne sont pas rigoureuses en raison de l'utilisation de nombres entiers pour W_i), la formule (2.3) peut être exprimée comme é

$$V_3(Y) = A \sum_n^I \pi_i \left(W_i \frac{\pi_i}{Y_i} - Y \right)^2 + W \sum_n^I \left(1 - \frac{M_i}{m_i} \right) M_i s_i^2. \quad (2.11)$$

(3) Estimateur de la variance de Rao (1975)

En ce qui concerne cet estimateur, on suppose que m_i unités secondaires sont prélevées suivant un EAS; toutefois, comme il s'agit d'un plan à auto-pondération, la taille m_i de l'échantillon au second degré est considérée comme une variable aléatoire. L'estimateur de la variance est défini par

$$V_4(Y) = A \sum_n^I \pi_i \left(W_i \frac{\pi_i}{Y_i} - Y \right)^2 + \sum_n^I \frac{\pi_i}{d_i^2} \left(\frac{\pi_i}{d_i^2} - A \right) \left(\frac{\pi_i}{d_i^2} - \frac{1}{Y} \right) \left\{ M_i^2 s_i^2 - \frac{m_i}{d_i} \sum_n^I \frac{\pi_i}{d_i} M_i s_i^2 \right\} \quad (2.12)$$

où A est défini en (2.4) et s_i^2 est défini en (2.5). Après quelques simplifications, nous pouvons réécrire (2.7) de la façon suivante:

$$V_4(Y) = V_3(Y) + W^2 \sum_n^I m_i s_i^2 \left\{ \left(1 - \frac{W}{m_i} \right) - A \left(\frac{\pi_i}{1} - 1 \right) \right\}. \quad (2.13)$$

Nous constatons que la formule de la variance comporte un terme additionnel, qui peut être positif ou négatif, lorsqu'on suppose que la taille de l'échantillon au second degré est aléatoire.

Comme W_{ij} est l'intervalle d'échantillonnage défini pour l'échantillonnage systématique ; logements dans la grappe prélevée dans le i -ième groupe aléatoire, on le définit comme un nombre entier pour simplifier les opérations.

Une UPE est prélevée avec probabilité proportionnelle à W_{ij} dans chacun des n groupes aléatoires. L'UPE j prélevée dans le i -ième groupe aléatoire est sous-échantillonnée systématiquement suivant un taux de sondage $1/W_{ij}$. Donc, le taux de sondage global dans chacun des n groupes aléatoires est $1/W$, de sorte que le plan devient un plan à auto-pondération avec un poids de base de W . On attribue à chaque groupe aléatoire un numéro de renouvellement de 1 à 6. Le nombre de groupes aléatoires n est habituellement un multiple de six de sorte qu'il y a le même nombre de groupes aléatoires pour chaque groupe de renouvellement. Comme une seule UPE est échantillonnée dans chaque groupe aléatoire, nous désignons par $1/W_i$ le taux de sous-échantillonnage dans l'UPE prélevée dans le i -ième groupe aléatoire et par m_i le nombre de logements échantillonnés du groupe aléatoire i .

2 Estimateurs de la variance

Supposons que nous voulons connaître le total d'une caractéristique y pour la strate. Soit y_{jk} la valeur de la caractéristique pour le k -ième logement de la j -ième UPE, $y_{jk} = \sum_{k=1}^{M_j} y_{jk}$. Nous pouvons alors estimer le total $Y = \sum_{j=1}^N \sum_{k=1}^{M_j} y_{jk}$ par $\bar{Y} = W \sum_{i=1}^n \bar{y}_i$, où \bar{y}_i est la somme des valeurs de la caractéristique y pour les m_i logements échantillonnés dans l'UPE qui a elle-même été prélevée dans le i -ième groupe, $i = 1, 2, \dots, n$. Nous considérons ci-dessous divers estimateurs de la variance du total estimé \bar{Y} :

(1) Estimateur de la variance de Keyfitz (1957)

Cet estimateur était utilisé dans l'ancien plan de sondage avec deux pseudo-échantillons répétés qui étaient formés, dans un cas, des groupes de renouvellement à numéro impair et dans l'autre cas, des groupes à numéro pair. Si on ne tient pas compte de la correction pour population finie (CPF), la formule de la variance est

(2.1)
$$V_1(\bar{Y}) = W^2 \left(\sum_{i=1}^n \bar{y}_i - \sum_{i=1}^n \bar{y}_i^2 \right)$$

où \sum désigne la sommation pour tous les groupes à numéro impair et \sum désigne la sommation pour tous les groupes à numéro pair. Par ailleurs, on peut utiliser l'estimateur de la variance Keyfitz généralisé pour $n(\geq 2)$ échantillons répétés, qui est défini par

(2.2)
$$V_2(\bar{Y}) = W^2 \frac{n-1}{n} \sum_{i=1}^n (\bar{y}_i - \bar{\bar{y}})^2$$

où $\bar{\bar{y}} = (1/n) \sum_{i=1}^n \bar{y}_i$. Dans ce cas, chaque UPE ou groupe de renouvellement est considéré comme un échantillon répété. On s'est intéressé à l'estimateur V_2 parce qu'on croyait qu'il pouvait être plus efficace (stable) que V_1 en raison du plus grand nombre de degrés de liberté.

comparés aux estimateurs de la méthode de Keyfitz à l'aide d'une simulation de Monte Carlo. Nous avons comparé le biais et la stabilité de ces divers estimateurs dans deux situations différentes: avec et sans correction par le quotient. Nous avons aussi étudié l'effet d'un accroissement du nombre d'échantillons répétés sur l'estimateur de la variance de Keyfitz. La section suivante donne un compte rendu de ces analyses. Compte tenu des résultats de l'analyse, nous avons étendu la méthode de Keyfitz aux secteurs AR.

Dans la section 3, nous établissons la formule de la variance de Keyfitz pour les estimations itératives par le quotient pour tous les genres de secteurs de l'EPA et analysons cette formule par une étude de Monte Carlo. Enfin, dans la section 4, nous tirons les conclusions qui s'imposent.

2. ESTIMATION DE LA VARIANCE POUR LE PLAN DE SONDAGE DES SECTEURS AR

2.1 Plan de sondage des secteurs AR

Pour les secteurs AR, le plan de sondage de l'EPA est un plan à deux degrés fondé sur la méthode des groupes aléatoires (Rao et coll. 1962) avec probabilité de sélection des unités primaires (UP) proportionnelle à la taille (PPT) et échantillonnage systématique des logements au second degré de telle sorte que le plan devient un plan à auto-pondération. Supposons qu'une strate donnée contient N UP et définissons respectivement x_j et M_j , $j = 1, 2, \dots, N$, comme la taille de la j -ième UP de la strate et son nombre de logements. Définissons $1/W$ comme le taux de sondage dans la strate, W étant un nombre entier, et n comme le nombre d'UP qui doivent être prélevées dans la strate. Les N UP contenues dans la strate sont réparties aléatoirement en n groupes de sorte que le i -ième groupe aléatoire contient N_i UP et $N'_i = N$.

Définissons

$$p_j = \frac{x_j}{N}, \quad j = 1, 2, \dots, N,$$

$$\sum_{i=1}^I x_i$$

et

$$\delta_{ij} = 1 \text{ si la } j\text{-ième UP est incluse dans le } i\text{-ième groupe}$$

$$= 0 \text{ dans le cas contraire.}$$

Alors, $\pi_i = \sum_{j=1}^N \delta_{ij} p_j$ est la taille relative du i -ième groupe aléatoire. Définissons maintenant W_{ij} , l'intervalle d'échantillonnage pour l'échantillonnage systématique. Posons $a_{ij} = \delta_{ij} W_{ij} / \pi_i$ et $r_{ij} = a_{ij} - [a_{ij}]$ où $[a]$ est le plus grand nombre entier égal ou inférieur à a . Sans limiter la généralité de ce qui précède, nous pouvons supposer que les éléments de l'ensemble $\{r_{ij}, j = 1, 2, \dots, N\}$ sont exprimés par ordre décroissant. Alors, W_{ij} est défini par

$$W_{ij} = [a_{ij}] + 1, \quad j = 1, 2, \dots, R$$

$$= [a_{ij}], \quad j = R + 1, \dots, N$$

où $R = \sum_{j=1}^N r_{ij}$. Puis, par définition, $\sum_{j=1}^N W_{ij} = W$ pour le i -ième groupe aléatoire, $i = 1, 2, \dots, n$.

Estimation de la variance pour l'enquête sur la population active du Canada

G.H. CHOUDHRY et H. LEE¹

RÉSUMÉ

Un moyen d'une étude de Monte Carlo, les auteurs évaluent le biais et la stabilité de divers estimateurs de la variance pour le plan de sondage à deux degrés fondé sur la méthode des groupes aléatoires (Rao et coll., 1962) dans le contexte de l'enquête sur la population active du Canada. Ils se servent de la méthode de linéarisation de Taylor pour déterminer la formule de la variance se rapportant à la technique d'estimation itérative par le quotient. Enfin, ils analysent les propriétés de cette formule par une simulation de Monte Carlo.

NOTES CLÉS: Estimateur de la variance de Keyfitz; estimateur itératif par le quotient; linéarisation de Taylor; simulation de Monte Carlo.

1. INTRODUCTION

L'enquête sur la population active du Canada (EPA), la plus vaste enquête mensuelle sur les ménages menée par Statistique Canada, sert à produire des estimations de diverses caractéristiques de la population active aux niveaux national, provincial et infraprovincial. Elle repose sur un plan de sondage stratifié à plusieurs degrés avec six groupes de renouvellement (Platek et Singh 1976).

L'échantillon de l'EPA est remanié après chaque recensement décennal de la population. Dans le cadre du remaniement entamé après le recensement de 1981, un programme de recherche intensif a été mis en oeuvre pour examiner diverses méthodes d'échantillonnage, d'estimation et de collecte des données (Singh et Drew 1981). La méthode d'estimation par le quotient pour domaines post-stratifiés, utilisée dans l'ancien plan, a fait place à la technique d'estimation itérative par le quotient afin d'accroître la fiabilité des données infraprovinciales. Dans cet article, nous nous intéressons plus particulièrement aux méthodes d'estimation de la variance.

La méthode d'estimation de la variance utilisée dans l'ancienne version de l'EPA reposait sur la généralisation de la méthode de Keyfitz (Keyfitz 1957) proposée par Woodruff Woodruff 1971), selon laquelle les estimations par le quotient pour les domaines post-stratifiés étaient soumises à une linéarisation de Taylor (Platek et Singh 1976). Nous désignerons cette méthode comme la méthode de Keyfitz (voir Platek et Singh 1976).

Le plan de sondage de l'EPA renferme trois genres de secteurs, soit les secteurs auto-représentatifs (AR), qui sont composés des grandes villes, les secteurs non auto-représentatifs (NAR), qui correspondent aux petits centres urbains et aux régions rurales, et les secteurs spéciaux, qui comprennent les établissements militaires, les institutions et les régions éloignées. En ce qui concerne les secteurs NAR et les secteurs spéciaux, nous avons choisi d'appliquer une version modifiée de la méthode de Keyfitz, qui fait intervenir l'estimation itérative par le quotient.

Par ailleurs, pour ce qui est des secteurs AR, nous avons analysé deux estimateurs de la variance définis respectivement par Rao, Hartley et Cochran (1962) et par Rao (1975) pour le plan de sondage à deux degrés fondé sur la méthode des groupes aléatoires et les avons

n^{-3} . Par conséquent, \hat{p} ne satisfait pas à l'hypothèse de Cornish-Fisher selon laquelle $\kappa_r(\hat{p}) = O(n^{1-r})$ pour $r \geq 1$: voir par exemple Kendall et Stuart (1977). On peut aussi déterminer les moments et les cumulants à l'aide de la f.g.m. (fonction génératrice des moments), que nous définissons ci-dessous.

$$E(\exp(t_1 M_1/n) | N_1) = \exp(t_1 N_1/n) S(t_1, n_1),$$

où

$$S(t_1, n_1) = (1 - n_1/l) \exp(-n_1 t_1/n) + (n_1/l) \exp((l - n_1) t_1/n).$$

Par (2.4), la f.g.m. est

$$E(\exp(t' \hat{p})) = E(\exp(t' N/n)) S(t) \text{ où } S(t) = \prod_1^r S(t_i, n_i).$$

De plus, à $t = 0$, $S_1 = 0$ et $S_{ij} = 0$ si l'indice inférieur ne se répète pas. Par exemple si nous posons

$$S = S(t), \quad \partial_i = \partial / \partial t_i, \quad S_i = \partial_i S, \quad S_{ij} = \partial_i \partial_j S, \quad \dots$$

nous obtenons

$$E(p_1^i \exp(t' \hat{p})) = E(\exp(t' N/n) [p_1^{i*} S + 2p_1^* S_1 + S_{11}]),$$

$$E(p_2^i \exp(t' \hat{p})) = E(\exp(t' N/n) [p_2^{i*} S + 2p_2^* S_2 + S_{22}] +$$

$$2p_1^* (p_2^{i*} S_1 + 2p_2^* S_{12} + S_{122}) + (p_2^{i*} S_{11} + 2p_2^* S_{112} + S_{1122})).$$

$$\text{Ainsi } E(p_2^i) = E(p_2^{i*} + S_{11}(0)) \text{ et}$$

$$E(p_2^i p_2^j) = E(p_2^{i*} p_2^{j*} + p_1^{i*} S_{22}(0) + p_2^{i*} S_{11}(0) + S_{1122}(0)).$$

où $S_{ii}(0) = S_{11}(0, n_i) = n^{-2} (l - n_i) n_i = n^{-2} \sum_{k=0}^{l-1} (l - k) k I(n_i = k)$ et $S_{1122}(0) = S_{11}(0) S_{22}(0)$. Nous pouvons simplifier davantage en utilisant $N_2 | N_1 \sim B(l, \theta, n - N_1)$ où $\theta = p_2 / (1 - p_1)$. La f.g.m. multinomiale donne

$$E(p_1^{i*} p_2^{j*}) = n^{-4} p_1 p_2 \{ (n)_4 p_1 p_2 + (n)_3 (p_1 + p_2) + (n)_2$$

$$\text{où } (n)_i = n! / (n - i)! = n(n - 1) \dots (n - i + 1).$$

BIBLIOGRAPHIE

- GASTWIRTH, J.L., KRIEGE, A.M., et RUBIN, D.B. (1978). Statistical analyses from summary data and their impact on the issue of confidentiality. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 183-188.
- KENDALL, M.G., et STUART, A. (1977). *The Advanced Theory of Statistics, Volume 7*. London: Griffin.
- NARGUNDAR, M.S., et SAVELAND, W. (1972). Random rounding to prevent statistical disclosures. *Proceedings of the Social Statistics Section, American Statistical Association*, 382-385.
- PENNY, R., et RYAN, M. (1986). A problem associated with random rounding. *New Zealand Statistician*, 21, 43-52.
- WITHERS, C.S. (1987a). Bias reduction by Taylor series. *Communications in Statistics - Theory and Methods* (en voie de rédaction).
- WITHERS, C.S. (1987b). Jackknifing binomials and multinomials. Document non-publié, Department of Scientific and Industrial Research.

Puisque $\Delta = 2/3f(0)$ à $p_1 = 0$, et est égale à $2/3$, 0 et $2/3$ pour (a), (b) et (c) respectivement. La convergence cesse à $p_1 = 0$ pour (a) et (c), mais non pour (b). Pour $n = 18$, on (a) et pour les valeurs de p_1 situées dans l'intervalle (.1, .8) pour les (b) et (c). Pour $n = 54$, ces intervalles deviennent (.1, .9) pour (a), (.02, .95) pour (b), et (.07, .95) pour (c). Les figures 2a et 2b décrivent la relation entre $Y = \log(-\log|\Delta|)$ et $X = \log(n)$ pour (a) $f(p) = 1$ et (b) $f(p) = p_1$. Comme prévu abstraction faite des petites valeurs de n , les courbes sont à peu près parallèles à la courbe $Y = X$ (sauf en ce qui concerne la fonction c) pour $p = .01$, ce qui est conforme à la relation $\Delta = O(e^{-\lambda n})$ pour certaine valeur de $\lambda > 0$. Les courbes ne sont pas lisses parce que Δ n'a été calculé qu'en fonction de n puissance 2 ($n = 2^i$ pour $0 \leq i \leq 7$). La figure 3 décrit la relation entre $Y = -\log|\Delta|$ et $X = \log(n)$ pour (c) $(p) = \exp(p_1)$. Pour les valeurs supérieures de n , les courbes sont parallèles à la courbe $Y = X$ pour $p_1 = .5$ et .1, ce qui est conforme à $\Delta = O(n^{-1})$, pour $p_1 = 0.1$ toutefois, la courbe ne reflète pas du tout une relation linéaire; le rythme d'accroissement est beaucoup plus prononcé. Les graphiques confirment de façon générale nos hypothèses sur la vitesse de convergence de (A.1). Si nous voulions vérifier ces hypothèses par des méthodes analytiques, il nous faudrait recourir à la théorie des nombres.

ANNEXE B

Dans cette annexe, nous comparons les moments et les cumulants de $\mathbf{p}^* = N/n$ et $\mathbf{p} = M/n$. Posons $q_1 = 1 - p_1$, $n_i = N_i \bmod l$, et $m_i(f) = E(p_i^l I(n_i = f)) \rightarrow p_i^l / l$ lorsque $n \rightarrow \infty$, en supposant que $p_1 \neq 0$ or 1.

Par des calculs élémentaires, nous obtenons

$$\begin{aligned} \mu(\mathbf{p}) &= \mu(\mathbf{p}^*) = \mathbf{p}, \\ \mu_2(\mathbf{p}_1) &= \mu_2(\mathbf{p}_1^*) + M_{22}n^{-2} = p_1q_1n^{-1} + O(n^{-2}), \\ M_{22} &= A^n(p_1) = \sum_{i=1}^l i!(l-i)m_0(i) \rightarrow (l^2 - 1)/6 \end{aligned}$$

orsque $n \rightarrow \infty$,

$$\begin{aligned} \mu_3(\mathbf{p}_1) &= \mu_3(\mathbf{p}_1^*) + 3n^{-2} \sum_{i=1}^l i!f^2(m_2(f) - 2p_1m_1(f) + p_1^2m_0(f)) \\ &+ n^{-3} \sum_{i=0}^f a_{jlm_0}(f) \end{aligned}$$

$$\begin{aligned} \mu_3(\mathbf{p}_1^*) &+ O(n^{-2}) = p_1q_1(1 - 2p_1)n^{-2} + O(n^{-2}), \\ a_{jlm_0}(f) &= -f^3(1 - f/l) + (l - f)^3j/l. \end{aligned}$$

De même $\mu_4(\mathbf{p}_1)$ peut s'écrire $\mu_4(\mathbf{p}_1^*) + \Sigma_2^f M_{4l}n^{-l} = O(n^{-2})$ et $\mu_4(\mathbf{p}_1)$ peut s'écrire $\Sigma_2^f k_4n^{-l}$ où $k_{42} = M_{42}$ ne tend pas vers 0 lorsque $n \rightarrow \infty$. D'où $k_4(\mathbf{p}_1) \sim n^{-2}$, et non

Insert Figure 2 (b), 3

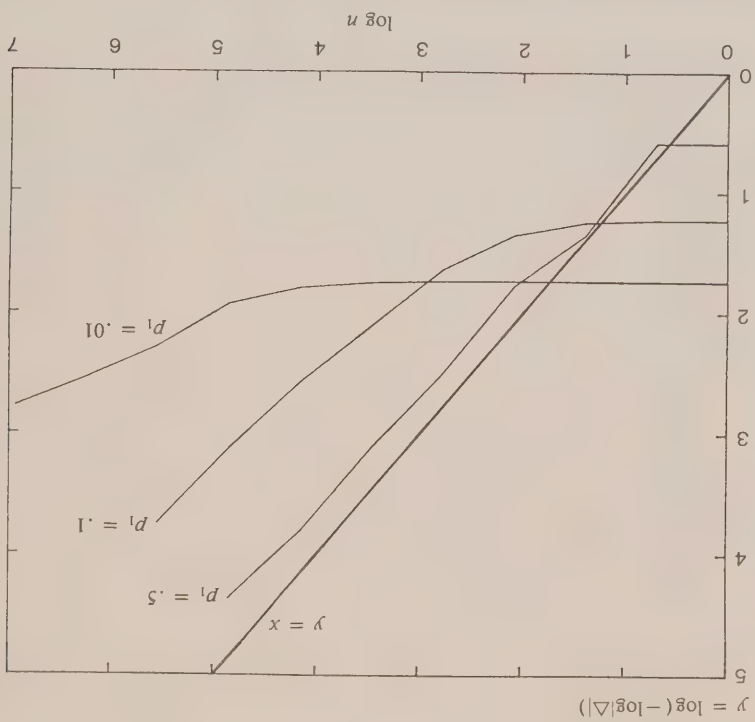


Figure 2b. Confirmation de la convergence exponentielle en (A.1) lorsque $f(p) = p_1$.

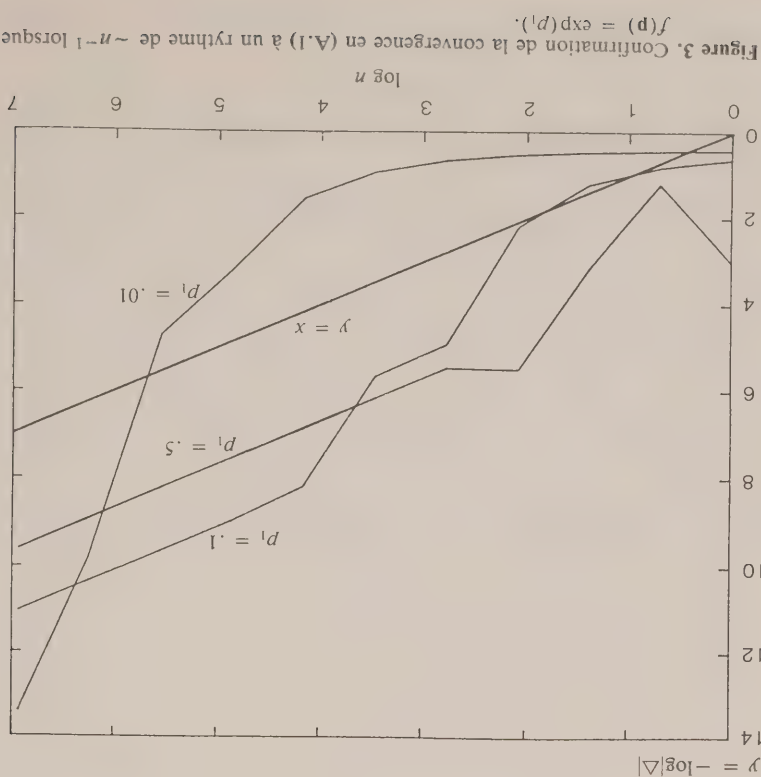


Figure 3. Confirmation de la convergence en (A.1) à un rythme de $\sim n^{-1}$ lorsque $f(p) = \exp(p_1)$.

Insert Figure 1(c), 2(a)

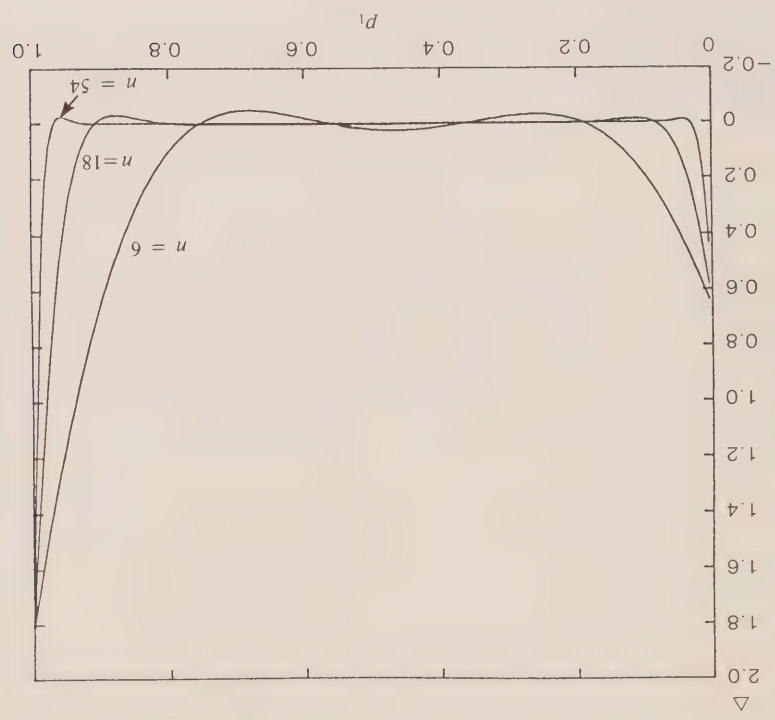


Figure 1c. Confirmation de l'hypothèse (A.1) lorsque $f(p) = \exp(p_1)$.

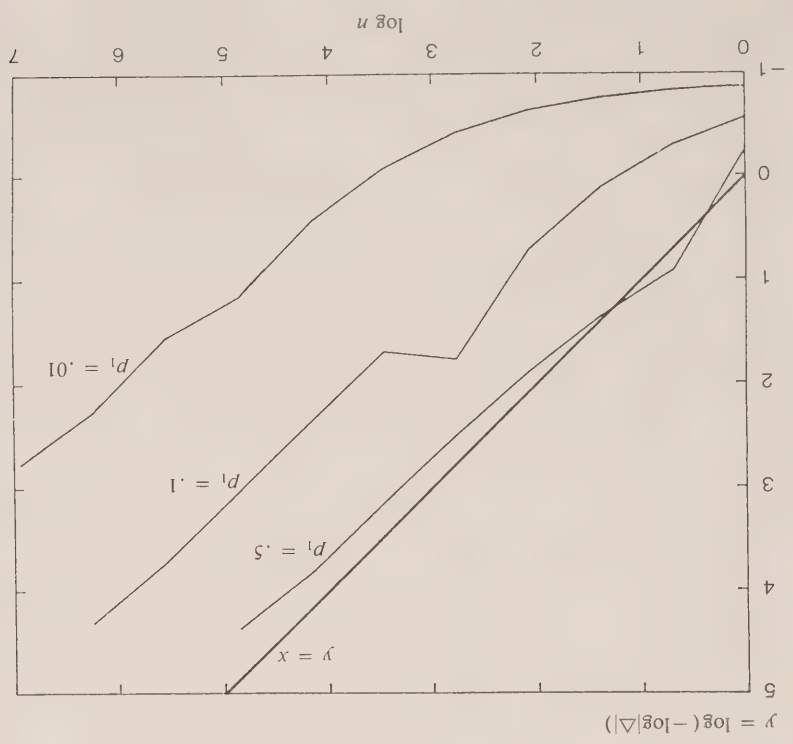


Figure 2a. Confirmation de la convergence exponentielle en (A.1) lorsque $f(p) = 1$.

Insert Figure 1 (a), 1(b)

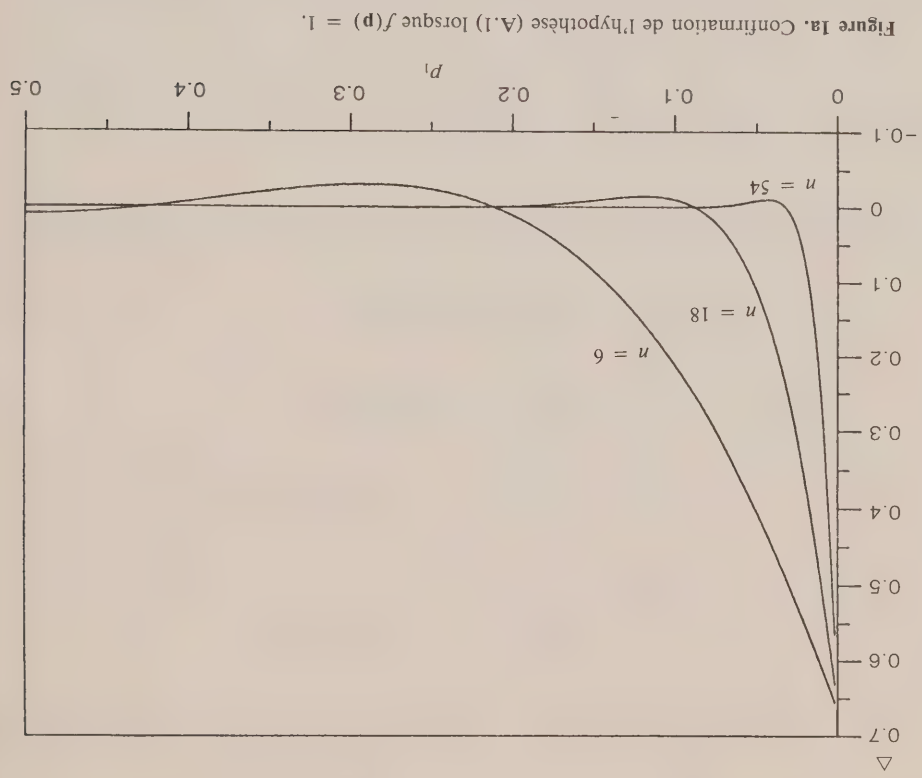


Figure 1a. Confirmation de l'hypothèse (A.1) lorsque $f(p) = 1$.

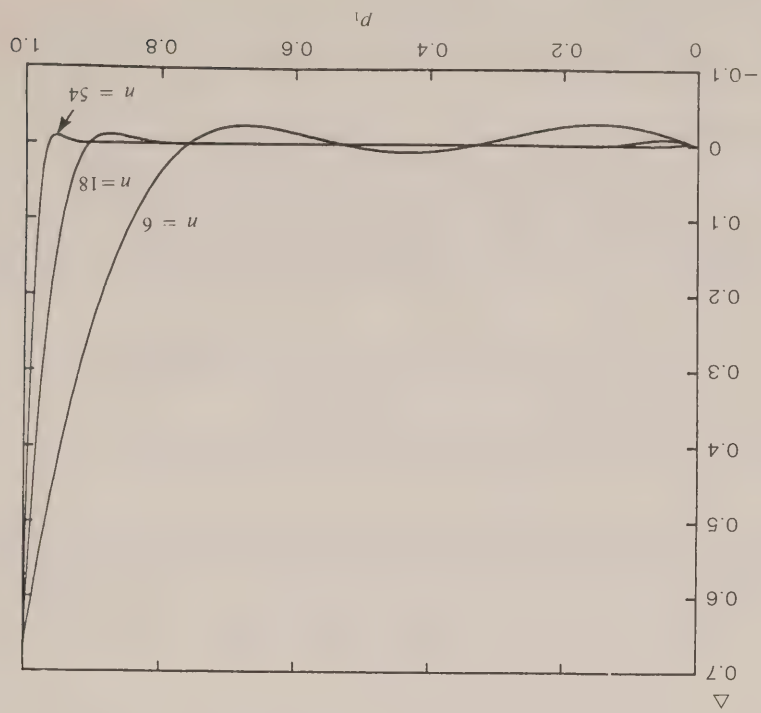


Figure 1b. Confirmation de l'hypothèse (A.1) lorsque $f(p) = p_1$.

Preuve du théorème 1.5. La première équation de l'expression (1.16) découle de (2.4) tandis que la seconde découle du théorème multinomial. Le corollaire 1.2 s'ensuit automatiquement.

Preuve du corollaire 1.3. Selon (1.16), pour $1 \leq l < i \leq R$ nous avons

$$E(f_l(\mathbf{p})p_i a_{n,l+1}) = f_l(\mathbf{p})p_i a_{n,l+1}$$

alors

$$E(f_l(\mathbf{p})) \sum_R^{l+1} p_l / a_{n,l+1} = f_l(\mathbf{p}) (1 - \sum_1^l p_i)$$

$$= E(f_l(\mathbf{p}) / a_{nl}) - f_l(\mathbf{p}) \sum_1^l p_i.$$

REMERCIEMENTS

Je tiens à remercier Peter McGavin pour les calculs de l'annexe A.

ANNEXE A

Pour une fonction lisse f on devrait avoir:

$$(A.1) \quad E(f(\mathbf{p})I(N_1 = j_1 \bmod l, \dots, N_s = j_s \bmod l)) \rightarrow f(\mathbf{p})l^{-s}$$

lorsque $n \rightarrow \infty$ pourvu que $0 < p_i < 1$ pour $1 \leq i \leq s \leq R$.

Si $E(f(\mathbf{p})) = f(\mathbf{p})$, on devrait observer une vitesse de convergence exponentielle, c'est-à-dire $O(e^{-\lambda n})$ pour certaine valeur de $\lambda > 0$. Si $f(\mathbf{p})$ est biaisé, le biais est alors $O(n^{-1})$, et cette valeur devrait représenter une vitesse de convergence pour (A.1). En règle générale, la convergence cesse lorsque \mathbf{p} approche les limites de $[0, 1]^r$, puisque

$$E(f(\mathbf{p})I(N_1 = j_1 \bmod l, \dots, N_s = j_s \bmod l))$$

$$= \begin{cases} f(\mathbf{p})I(j_1 = j_2 = \dots = j_s = 0) & \text{si } \mathbf{p} = 0 \\ f(\mathbf{p})I(j_1 = n \bmod l) & \text{si } p_1 = 1. \end{cases}$$

Pour vérifier ces hypothèses, nous avons considéré le cas où $s = 1$, $l = 3$, $j = 0$ et les fonctions (a) $f(\mathbf{p}) = 1$, (b) $f(\mathbf{p}) = p_1$, et (c) $f(\mathbf{p}) = \exp(p_1)$. Les calculs ont été effectués à quadruple précision sur un VAX11/780, ce qui a permis de calculer

$$\Delta = E(f(\mathbf{p})I(N_1 = j_1 \bmod l, \dots, N_s = j_s \bmod l)) - f(\mathbf{p})l^{-s}$$

à une précision de 112 bits, soit près de 34 décimales. Les figures 1a, 1b et 1c décrivent la relation entre Δ et p_1 pour $n = 6, 18, 54$. Puisque $n \bmod 3 = 0$, Δ est symétrique autour de $p_1 = 1/2$ pour (a).

puisque

$$(2.2) \qquad \sum_{l=1}^{l-1} l!(l-l) = (l^2 - 1)/6.$$

Or,

$$(2.3) \qquad E(p_1^{*2}) = p_1^2 + (p_1 - p_2^2)n^{-1},$$

ce qui nous amène à l'équation (1.5). Par ailleurs, $p_1 = \bar{p}_1 = \sum (M_j - N_j)/n$,

alors

$$E(\bar{p}_1^2) = E(\bar{p}_2^2) - 2n^{-2} \sum E(M_1(M_j - N_j) + n^{-2} \sum E((M_i - N_i)(M_j - N_j))) \\ = E(\bar{p}_2^2) - 2n^{-2} \sum A_n(p_1) + n^{-2} \sum A_n(p_i)$$

puisque $E(\Pi_i f_i(M_i) | \{N_j\}) = \Pi_i E(f_i(M_i) | N_i)$.

(2.4) Par conséquent $\text{var}(\bar{p}_1) = (p_1 - p_2^2)n^{-1} + n^{-2} \sum_{i \neq 1} A_n(p_i)$ ce qui vérifie l'équation (1.7).

Par ailleurs

$$E(\bar{p}_1 \bar{p}_i) = p_1 - n^{-2} \sum_{i \neq 1} E(M_1 M_i) = p_1 - \sum_{i \neq 1} E(p_1 p_i^*)$$

$$= p_1 - \sum_{i \neq 1} p_1 p_i (1 - n^{-1}) = p_1 - p_1 (1 - p_1) (1 - n^{-1}),$$

alors,

$$(2.5) \qquad \text{cov}(\bar{p}_1, \bar{p}_i) = (p_1 - p_2^2)n^{-1}.$$

Par conséquent, $\text{var}(p_1(\lambda)) = (p_1 - p_2^2)n^{-1} + \{(1 - \lambda)^2 A_n(p_1) + \lambda^2 \sum_{i \neq 1} A_n(p_i)\}n^{-1}$ ce qui vérifie (1.8).

Preuve du théorème 1.2. Cette démonstration a été faite pour \mathbf{p}^* dans Withers (1987a) En outre, comme f est finie dans un voisinage de \mathbf{p} ,

$$f(\hat{\mathbf{p}}) = f(\mathbf{p}^*) + (\hat{\mathbf{p}} - \mathbf{p}^*)' f'(\mathbf{p}^*) + O(|\hat{\mathbf{p}} - \mathbf{p}^*|^2).$$

$$E((\hat{\mathbf{p}} - \mathbf{p}^*) | N) = 0, E((\bar{p}_1 - p_1^*)^2 | N) = 2n^{-2} I(N_1 \bmod l \neq 0),$$

où $I(A) = 1$ ou 0 selon que A est vrai ou faux, autrement dit, $I(\cdot)$ est la fonction indicatrice. Par conséquent $E(f(\hat{\mathbf{p}})) = E(f(\mathbf{p}^*)) + O(n^{-2})$ et $\text{var}(f(\hat{\mathbf{p}})) = \text{var}(f(\mathbf{p}^*)) + O(n^{-2})$.

Preuve du théorème 1.3. Découle directement de (2.1) et de (2.3).

Preuve du théorème 1.4. Découle de (2.1) et de (1.5).

(1.17)

$$E(f_I(\mathbf{p})) = E(f_I(\mathbf{p}^*)) = f_I(\mathbf{p})a_{nI}.$$

$\mathbf{p})/a_{nI}$ est donc un estimateur sans biais de $f(\mathbf{p})$.

Corollaire 1.2. $\text{cov}(\hat{p}_1, \hat{p}_2) = -p_1 p_2 / n$. L'estimateur sans biais correspondant est $\hat{p}_1 \hat{p}_2 / (n - 1)$. D'une manière plus générale, pour $1 \leq I \leq R$, $E(\prod_{i=1}^I (\hat{p}_i - p_i)) = \prod_{i=1}^I p_i$ avec l'estimateur sans biais $(\prod_{i=1}^I \hat{p}_i) a_{nI} / c_{nI}$ où $c_{nI} = \sum_{f=0}^{I-1} (-1)^{I-f} \binom{I}{f} a_{nf}$. (On tient le même résultat si on remplace \mathbf{p} par \mathbf{p}^* .)

L'équation (1.16) nous permet de calculer des estimateurs sans biais pour d'autres lymphomes spéciaux en \mathbf{p} tels que $p_1^2, p_1 p_2 (p_1 + p_2)$ et $\sum_{i=1}^R p_i^3$ - mais non pour des polymes de la forme $p_1^2 p_2$ ou p_1^3 .

Corollaire 1.3. Pour $1 \leq I < R$ un estimateur sans biais de

$$(1.18) \quad f_I(\mathbf{p}) \sum_I p_i \text{ est } f_I(\mathbf{p}_1) \left\{ 1 - In^{-1} - \sum_R^{I+1} \hat{p}_i \right\} / a_{n,I+1}.$$

particulier, $\hat{p}_1 (\hat{p}_1 - n^{-1}) (1 - n^{-1})^{-1}$ est un estimateur sans biais de p_1^2 . (1.19)

Nous tenons à souligner que les résultats de cette étude reposent sur l'hypothèse que les séquences de tableaux sont des variables de Poisson indépendantes ou à tout le moins, des variables multinomiales, étant donné le total. Le modèle de Poisson et le modèle multinomial ont tous deux intéressants parce qu'ils sont d'interprétation simple et parce que la somme de variables de Poisson est elle-même une variable de Poisson. Cependant, la somme de variables multinomiales n'est multinomiale que si les probabilités de cellule \mathbf{p} sont les mêmes sur toutes les variables. Cela donne à penser que les modèles multinomiaux peuvent produire des conclusions moins exactes si les populations étudiées sont constituées de deux groupes non-homogènes ou plus.

2. PREUVES

Preuve du théorème 2.1. Posons $r = N_1 \bmod l$. Alors l'équation (1.1) est vraie pour $= N_1, M = M_1$ avec $jl = N - r$ et

$$E(M_1^r | r) = (N_1 - r)^2 (1 - r/l) + (N_1 - r + l)^2 r/l = N_1^2 + lr - r^2.$$

tr conséquent,

$$(2.1) \quad E(\hat{p}_1^2) = E(p_1^2) + n^{-2} A_n(p_1),$$

$$A_n(p_1) = E(M_1^2 - N_1^2) = E(lr - r^2) = \sum_{l=1}^l (ll - l^2) P(N = l)$$

$$= (l^2 - 1)/6 + \Delta_n(p_1)$$

(b) l'arrondissement aléatoire n'a qu'un effet du second ordre sur le biais de l'estimateur naturel de $f(\mathbf{p})$.
 Selon (1.12), l'estimateur naturel de $f(\mathbf{p})$, $f(\hat{\mathbf{p}})$, a un biais d'ordre n^{-1} . Nous allons voir maintenant comment réduire ce biais à un ordre n^{-2} .

Corollaire 1.1. Si pour une fonction $f_n(\mathbf{p})$, $E(f_n(\mathbf{p}^*)) = f(\mathbf{p}) + O(n^{-2})$, alors $E(f_n(\hat{\mathbf{p}})) = f(\mathbf{p}) + O(n^{-2})$.

Deux estimateurs possibles pour $f_n(\hat{\mathbf{p}})$ dont "l'estimateur delta" pour lequel

$$(1.1) \quad f_n(\mathbf{p}) = f(\mathbf{p}) - \left\{ \sum_{i=1}^I f_n(\mathbf{p}) p_i - \mathbf{p}' f(\mathbf{p}) \mathbf{p} \right\} / (2n),$$

où $f_n(\mathbf{p}) = \partial^2 f(\mathbf{p}) / \partial p_i^2$ et "l'estimateur jack-knife" pour lequel

$$(1.1) \quad f_n(\mathbf{p}) = n f(\mathbf{p}) - (n-1) f,$$

où $f = \sum_{i=1}^I p_i f([(n\mathbf{p} - e_i) / (n-1)]) + (1 - \sum_{i=1}^I p_i) f([n\mathbf{p} / (n-1)])$,

e_i = le i -ième vecteur unitaire dans R ,

$$\text{et } [x] : R' \rightarrow R' \text{ est définie par } [x]_i = \begin{cases} 0, & x_i < 0 \\ x_i, & 0 \leq x_i \leq 1 \\ 1, & x_i > 1 \end{cases}$$

Ces estimateurs ont été calculés dans Withers (1987a et 1987b). En particulier, si $f(\mathbf{p})$ est une fonction de p_1 , par exemple $f(\mathbf{p}) = g(p_1)$, alors $f_n(\mathbf{p}) = g(p_1) - \hat{g}(p_1)$ et $f = p_1 g([(np_1 - 1) / (n-1)]) + (1 - p_1) g([np_1 / (n-1)])$.
 Si par exemple, $f(\mathbf{p}) = p_1^2$ nous avons, pour l'estimateur delta, l'équation suivante

Nous allons maintenant montrer que si $f(\mathbf{p})$ est une fonction polynomiale, il est effectivement possible de trouver un estimateur de $f(\mathbf{p})$ fondé sur l'estimateur naturel avec un biais apparemment de taille exponentiellement faible. Pour cela, nous prenons le cas

$$f(\mathbf{p}) = p_1^2.$$

Théorème 1.3. $\lambda_1 = \{ p_1^2 - n^{-1} \hat{p}_1 - n^{-2} (l^2 - 1) / 6 \} (1 - n^{-1})^{-1}$ est un estimateur de $\lambda_1 = p_1^2$ avec un biais $\Delta_n(p_1) (n^2 - n)^{-1}$.

De même si $f_n(\mathbf{p})$ est un moment de $\hat{\mathbf{p}}$ il est possible de trouver un estimateur de $f_n(\mathbf{p})$ avec un biais apparemment de taille exponentiellement faible. À des fins d'illustration, nous prenons le cas $f_n(\mathbf{p}) = \text{var}(\hat{p}_1)$.

Théorème 1.4. $\lambda_{2n} = n^{-1} (\hat{p}_1 - \lambda_1) - n^{-2} (l^2 - 1) / 6$ est un estimateur de $\lambda_{2n} = \text{var}(\hat{p}_1)$ avec un biais $-\Delta_n(p_1) (n^2 - n)^{-1}$.

On peut étendre ces résultats à des polynômes d'ordre supérieur au moyen de l'expression définie dans l'annexe B pour les moments et les cumulants de $\hat{\mathbf{p}}$. Nous allons maintenant montrer qu'il existe un estimateur sans biais pour le cas particulier où $f(\mathbf{p})$ est collinéaire.

Théorème 1.5. Soit $f_I(\mathbf{p}) = \prod_{i=1}^I p_i$ où $1 \leq I \leq R$ et

$$a_{nI} = n^{-I} n! / (n - I)! = (1 - n^{-1}) (1 - 2n^{-1}) \dots (1 - (I - 1)n^{-1}). \quad (1.16)$$

Dans l'annexe A, nous démontrons que pour $0 < p_1 < 1$, $P(N_1 \bmod l = i) - 1 \rightarrow 0$ de façon exponentielle lorsque $n \rightarrow \infty$, de sorte que $\Delta_n(p_1) \rightarrow 0$ de façon exponentielle lorsque $n \rightarrow \infty$, et $\nabla_n(p)$ fait de même à la condition que $p_i \neq 0$ pour tous les i . Comme $\alpha(\lambda)$ est minimisée par $\lambda_R = R^{-1}$ et $\alpha(\lambda_R) = 1 - R^{-1}$, $\text{var}(p_1(\lambda))$ l'est aussi symptotiquement. Ainsi, la perte d'efficacité causée par l'utilisation de l'estimateur naturel au lieu de l'estimateur asymptotiquement optimal sans biais $p_1(\lambda_R)$ lorsque R est grand) est donnée par

$$\{\text{var}(\hat{p}_1) - \text{var}(p_1(\lambda_R))\} / \{\text{var}(p_1(\lambda_R))\} \approx (l^2 - 1) / \{6Rn(p_1 - p_1^2)\}, \quad (1.11)$$

quelque valeur est négligeable si $M_1(1 - M_1/n) \approx n(p_1 - p_1^2)$ est élevée par rapport à $2 - 1 / \{6R\}$.

En règle générale, M_1 est une bonne approximation du membre gauche de (1.11). Nous avons ainsi une façon pratique de vérifier l'efficacité des estimateurs naturels. Si une ou plusieurs valeurs $\{p_i\}$ sont nulles, $p_i = 0$, alors $N_i = M_i = 0$, nous disons que $\Sigma_{i \neq 1}$ ne comprend pas les cellules pour lesquelles $p_i = 0$, et que R désigne le nombre de cellules du tableau pour lesquelles $p_i \neq 0$.)

En nous servant de l'équation (1.5) nous pouvons maintenant faire une brève comparaison l'arrondissement aléatoire et de la contamination aléatoire. Selon nos informations, les organismes statistiques de l'Australie et du Royaume-Uni arrondissent les fréquences de cellule leur ajoutant 1 avec une probabilité 1/4, 0 avec une probabilité 1/2 et -1 avec une probabilité 1/4, de sorte que

$$\text{var}(\hat{p}_1) = (p_1 - p_1^2)n^{-1} + 1/2n^{-2}.$$

coefficient 1/2 passe à 4/3 dans le cas de la Nouvelle-Zélande ($l = 3$) et à 4 pour le Canada ($l = 5$). La méthode de la contamination aléatoire implique une moins grande protection ne variation maximale de 1 par opposition à 2 pour le système de Nouvelle-Zélande et à 3 pour le système canadien) et la possibilité d'une fréquence négative si elle est appliquée des cellules vides.

Le théorème 1.1 montre que l'arrondissement aléatoire n'a qu'un effet de second ordre sur l'efficacité de l'estimateur de p_1 ; la variance n'est accrue que par un facteur de l'ordre n^{-2} . La démonstration suivante montre que ce résultat très important est aussi vrai pour l'estimation de n importe quelle fonction lisse de $\{p_i\}$. Posons $r = R - 1$, $\mathbf{p} = (p_1, \dots, p_r)$, $\mathbf{p}^* = (N_1, \dots, N_r)$, $\mathbf{M} = (M_1, \dots, M_r)$, $\mathbf{p}^* = \mathbf{N}/n$ et $\hat{\mathbf{p}} = \mathbf{M}/n$. Alors nous avons $\mathbf{p}^*(\mathbf{p}^*) = V/n$ où $V = \text{diag}(\mathbf{p}^*)$. Supposons maintenant que nous voulons estimer \mathbf{p} , une fonction dont les secondes dérivées sont continues. Ainsi, $f(\mathbf{p}) = \partial f(\mathbf{p}) / \partial \mathbf{p}$ est une fonction continue $r \times 1$ et $f(\mathbf{p}) = \partial^2 f(\mathbf{p}) / \partial \mathbf{p} \partial \mathbf{p}'$ est une fonction continue $r \times r$.

Théorème 1.2. Lorsque $n \rightarrow \infty$, $E(f(\mathbf{p}^*))$ et $E(f(\hat{\mathbf{p}}))$ sont toutes deux égales à

$$f(\mathbf{p}) + B(\mathbf{p})n^{-1} + O(n^{-2}) \text{ où } B(\mathbf{p}) = \text{trace}(f'(\mathbf{p})V/2). \quad (1.12)$$

et plus, $\text{var}(f(\mathbf{p}^*))$ et $\text{var}(f(\hat{\mathbf{p}}))$ sont toutes deux égales à

$$v(\mathbf{p})n^{-1} + O(n^{-2}) \text{ où } v(\mathbf{p}) = f'(\mathbf{p})'Vf(\mathbf{p}). \quad (1.13)$$

et le théorème montre que

(l'arrondissement aléatoire n'accroît que de $O(n^{-2})$ la variance de l'estimateur naturel de $f(\mathbf{p})$).

de ces méthodes, voir Gastwirth et coll. (1978) et Kendall et Stuart (1977). On trouvera également une bibliographie sur l'arrondissement aléatoire pour les données multidimensionnelles et les données groupées dans Gastwirth et coll. (1978).

Dans cet article, nous allons nous concentrer sur les problèmes liés à l'estimation d'une fonction de probabilités de cellule associées à un tableau de R valeurs arrondies aléatoirement. Pour des raisons de commodité, nous désignons les probabilités de cellule par p_1, \dots, p_R au lieu de $\{p_j, 1 \leq i \leq I, 1 \leq j \leq J\}$, comme cela se fait normalement pour un tableau $I \times J$.

Ainsi, $1 = \sum_R^1 p_i$ et $n = \sum_R^1 N_i$ est la somme des fréquences contenues dans le tableau. Soient $\{M_j\}$ les valeurs arrondies de $\{N_j\}$. Etant donné n , nous supposons que $\{N_j\}$ suit une loi multinomiale avec les paramètres n et $\{p_j\}$. L'hypothèse ci-dessus est vraie pour $p_i = m_i / \sum_j m_j$ si $\{N_j\}$ sont non conditionnellement des variables de Poisson indépendantes avec des moyennes $\{m_j\}$.

Deux estimateurs sans biais de p_1 sont définis par:

(1.2)
$$p_1^* = N_1/n \text{ et } \hat{p}_1 = M_1/n.$$

Le premier n'est pas un vrai estimateur puisqu'on ne connaît pas N_1 . Le second est l'estimateur naturel. (Nous supposons que n est connu. S'il ne l'est pas nous pouvons sans difficulté le remplacer par $\sum_R^1 M_i$.) Il existe toutefois d'autres estimateurs sans biais, notamment "l'estimateur complémentaire"

(1.3)
$$\hat{p}_1 = - \sum_{j \neq 1} M_j/n,$$

d'où

(1.4)
$$p_1(\lambda) = (1 - \lambda)\hat{p}_1 + \lambda \hat{p}_1 \text{ pour n'importe valeur } \lambda \text{ donnée.}$$

Il faut alors se demander quelle est la meilleure valeur de λ que l'on peut utiliser et quelle perte d'efficacité doit-on subir si l'on utilise l'estimateur naturel c'est-à-dire $\lambda = 0$. Pour répondre à ces questions, il faut connaître les variances de ces estimateurs. Elles sont définies ci-dessous.

Théorème 1.1

Soit

(1.5)
$$\text{var}(\hat{p}_1) = (p_1 - p_2^2) n^{-1} + \{ (l^2 - 1) / 6 + \Delta_n(p_1) \} n^{-2} = v_n(p_1),$$

où

(1.6)
$$\Delta_n(p_1) = \sum_{l=1}^I l(l-1) \{ P(N_l \text{ mod } l = l) - l^{-1} \}.$$

Aussi,

(1.7)
$$\text{var}(\hat{p}_1) = (p_1 - p_2^2) n^{-1} + \{ (R - 1) (l^2 - 1) / 6 + \sum_{j \neq 1} \Delta_n(p_j) \} n^{-2},$$

et

(1.8)
$$\text{var}(p_1(\lambda)) = (p_1 - p_2^2) n^{-1} + \{ \alpha(\lambda) (l^2 - 1) / 6 + \Delta_n(p) \} n^{-2},$$

où

(1.9)
$$\alpha(\lambda) = (1 - \lambda)^2 + (R - 1) \lambda^2$$

et

(1.10)
$$\Delta_n(p) = (1 - \lambda)^2 \Delta_n(p_1) + \lambda^2 \sum_{i \neq 1} \Delta_n(p_i).$$

(Les démonstrations se trouvent à la section 2.)

Tableau 4
Rendement de divers estimateurs selon le modèle de réponse B

estimateurs	TC	TI	TD ₁	TD ₃	TD ₅	TD ₁ *	TD ₃ *	TD ₅ *
Taux de réponse moyen $\bar{q} = .60$								
t = 50	BIAS	.015	1.086	.290	.383	.716	.323	.688
	VAR	.405	.966	1.208	1.011	.937	1.050	.907
	EQM	.405	2.145	1.29	1.158	1.450	1.154	1.380
t = 100	BIAS	.007	1.079	.120	.349	.732	.196	.668
	VAR	.186	.422	.513	.429	.420	.447	.401
	EQM	.186	1.586	.527	.551	.956	.485	.847
POP2								
t = 50	BIAS	.090	4.046	1.362	1.757	2.826	1.562	2.749
	VAR	3.952	10.285	12.519	12.089	12.010	11.605	11.046
	EQM	3.960	26.655	14.374	15.176	19.996	14.045	18.603
t = 100	BIAS	.056	3.897	.454	1.531	2.707	.853	2.521
	VAR	1.710	4.151	5.432	5.121	5.103	4.798	4.541
	EQM	1.713	19.338	5.638	7.465	12.431	5.525	10.896
Taux de réponse moyen $\bar{q} = .70$								
t = 50	BIAS	.015	.584	.179	.221	.409	.196	.376
	VAR	.405	.751	.826	.425	.716	.769	.723
	EQM	.405	1.092	.858	.474	.883	.807	.864
t = 100	BIAS	.007	.536	.046	.173	.365	.087	.317
	VAR	.186	.307	.318	.295	.295	.299	.295
	EQM	.186	.594	.320	.325	.428	.307	.395
POP2								
n = 50	BIAS	.090	2.057	.682	.891	1.477	.804	1.392
	VAR	3.952	6.199	6.788	6.165	6.232	6.340	6.093
	EQM	3.960	10.430	7.253	6.959	8.414	6.986	8.031
n = 100	BIAS	.056	1.918	.157	.755	1.311	.374	1.175
	VAR	1.710	2.826	2.897	2.884	2.867	2.796	2.836
	EQM	1.713	6.506	2.922	3.454	4.586	2.936	4.217
POP1								
n = 50	BIAS	.015	.584	.179	.221	.409	.196	.376
	VAR	.405	.751	.826	.425	.716	.769	.723
	EQM	.405	1.092	.858	.474	.883	.807	.864
n = 100	BIAS	.007	.536	.046	.173	.365	.087	.317
	VAR	.186	.307	.318	.295	.295	.299	.295
	EQM	.186	.594	.320	.325	.428	.307	.395
POP2								
n = 50	BIAS	.090	2.057	.682	.891	1.477	.804	1.392
	VAR	3.952	6.199	6.788	6.165	6.232	6.340	6.093
	EQM	3.960	10.430	7.253	6.959	8.414	6.986	8.031
n = 100	BIAS	.056	1.918	.157	.755	1.311	.374	1.175
	VAR	1.710	2.826	2.897	2.884	2.867	2.796	2.836
	EQM	1.713	6.506	2.922	3.454	4.586	2.936	4.217

la stabilité de l'estimateur corrigé. Nous avons pu établir que la valeur optimum de h se situait autour de 0.1 pour notre analyse. Des résultats qui sont tirés de la même simulation mais qui ne sont pas reproduits ici indiquent que si l'on réduit davantage la valeur de h , le biais tend à s'accroître. Ce résultat est prévisible puisque en faisant tendre h vers 0, on obtient une série d'estimations \hat{q}_k ($k = 1, \dots, n$) égales à 1 ou à 0 selon qu'il s'agit de répondants ou de non-répondants.

6. REMERCIEMENTS

Je tiens à exprimer ma reconnaissance au professeur Luigi Biggieri pour le soutien qu'il m'a accordé au cours de la préparation de cet article. Je veux aussi remercier les arbitres pour les commentaires utiles qu'ils ont exprimés sur la version préliminaire.

6. Les estimateurs corrigés par l'estimation de la PRI ne sont pas très stables mais doivent être utilisés de préférence à TI si l'on se fonde sur l'EQM.

7. Comme prévu, le biais dépend directement du taux de non-réponse et de la différence entre le vrai modèle de superpopulation et le modèle hypothétique (c'est-à-dire, le "faux" modèle sur lequel reposent les estimateurs). L'utilisation de deux modèles de réponse (A et B) n'a pas d'incidence notable sur les résultats (voir Giommi (1984) pour les effets de divers modèles).

8. L'élargissement de l'échantillon semble se traduire par une légère diminution du biais pour tous les estimateurs étudiés. TD et TD^* sont des exceptions: dans ce cas, la réduction du biais n'est pas attribuable à la modification des conditions expérimentales mais à l'amélioration réelle de l'estimation q_k lorsque n augmente.

En conclusion, nous pouvons affirmer que les deux méthodes proposées dans cet article peuvent servir dans des situations comparables à celles que nous venons d'analyser et qu'il convient d'accorder une certaine préférence à la méthode (I) en raison de son application plus simple. Toutefois, nous n'avons pas résolu le problème de la détermination de la meilleure

Tableau 3

Rendement de divers estimateurs selon le modèle de réponse A

Estimateurs	TC	TI	TD_1	TD_3	TD_5	TD_1^*	TD_3^*	TD_5^*
-------------	------	------	--------	--------	--------	----------	----------	----------

$n = 50$	BIAS	.015	.861	.349	.420	.669	.380	.620
	VAR	.405	.973	1.115	1.036	1.007	1.041	.989
	EQM	.405	1.714	1.237	1.212	1.455	1.185	1.379
	BIAS	.007	.805	.164	.323	.610	.227	.544
$n = 100$	VAR	.186	.416	.443	.429	.412	.415	.404
	EQM	.186	1.064	.470	.533	.784	.467	.700
	BIAS	.007	.805	.164	.323	.610	.227	.544
	VAR	.186	.416	.443	.429	.412	.415	.404

POP2

$n = 50$	BIAS	.090	3.125	1.433	1.682	2.544	1.544	2.378
	VAR	3.952	8.744	9.821	9.823	9.743	9.390	9.233
	EQM	3.960	18.510	11.874	12.652	16.215	11.774	14.888
	BIAS	.056	2.959	.749	1.387	2.337	1.004	2.104
$n = 100$	VAR	1.710	4.144	4.515	5.122	4.819	4.238	4.632
	EQM	1.713	12.900	5.076	7.046	10.281	5.246	9.059
	BIAS	.056	2.959	.749	1.387	2.337	1.004	2.104
	VAR	1.710	4.144	4.515	5.122	4.819	4.238	4.632

Taux de réponse moyen $\bar{q} = .70$

POP1

$n = 50$	BIAS	.015	.581	.226	.271	.418	.249	.415
	VAR	.405	.765	.794	.750	.738	.754	.752
	EQM	.405	1.103	.845	.823	.913	.816	.924
	BIAS	.007	.531	.099	.205	.396	.143	.357
$n = 100$	VAR	.186	.328	.323	.307	.327	.313	.327
	EQM	.186	.610	.333	.349	.484	.333	.454
	BIAS	.007	.531	.099	.205	.396	.143	.357
	VAR	.186	.328	.323	.307	.327	.313	.327

POP2

$n = 50$	BIAS	.090	2.130	.813	.939	1.542	.887	1.453
	VAR	3.952	6.996	7.122	6.827	6.991	6.708	6.753
	EQM	3.960	11.533	7.783	7.709	9.396	7.495	8.864
	BIAS	.056	1.966	.473	.953	1.541	.658	1.406
$n = 100$	VAR	1.710	3.071	3.005	3.062	3.027	2.926	3.008
	EQM	1.713	6.937	3.229	3.970	5.402	3.359	4.985
	BIAS	.056	1.966	.473	.953	1.541	.658	1.406
	VAR	1.710	3.071	3.005	3.062	3.027	2.926	3.008

Tableau 2
Propriétés des populations simulées

Population	POP1			POP2		
	Moyenne	E.T.	C.V.	Moyenne	E.T.	C.V.
strate n° 1						
x	19.305	12.71	.66	20.037	14.50	.72
y	7.612	5.38	.71	1.961	2.21	1.13
strate n° 2						
x	50.325	21.32	.42	49.775	23.28	.47
y	30.325	13.38	.44	44.862	21.31	.47
Total						
x	34.815	23.42	.67	34.906	24.44	.70
y	18.969	15.26	.80	23.411	26.25	1.12

E.T. = Ecart type de la population; Asym. = asymétrie (3-ième moment/(2-ième moment)^{3/2}); C.V. = coefficient de variation.

où les paramètres θ , θ_1 , θ_2 sont choisis de telle manière que le taux de réponse moyen \bar{q} pour l'ensemble de la population soit de 0.6 ou de 0.7. En pratique, on forme des ensembles de répondants en exécutant un tirage de Bernoulli pour chaque unité $k \in s$ avec une probabilité de résultat favorable q_k (réponse) et une probabilité de résultat défavorable $1 - q_k$ (non réponse);

3) on estime la PRI à l'aide des méthodes (1) et (2) et on calcule pour chaque échantillon des valeurs de TC , TI , TD , TD^* ;

4) on répète 1000 fois les étapes 1 à 3 et on calcule à la fin le biais, la variance (VAR) et l'erreur quadratique moyenne (EQM) des estimateurs pour chaque combinaison possible des éléments suivants: taille de l'échantillon (50, 100), modèle de réponse (A, B), taux de réponse moyen (0.6, 0.7) et population (POP1, POP2).

Les résultats de la simulation figurent dans les tableaux 3 et 4.

5. RÉSULTATS DE L'ÉTUDE DE MONTE CARLO

L'analyse des tableaux 2 et 3 fait ressortir des points intéressants.

1. Comme prévu, TC est approximativement non biaisé dans toutes les simulations.
2. Le biais de TI est constamment supérieur à celui de TD et de TD^* . Il convient donc, du moins dans les situations qui s'apparentent à celles de la simulation, d'utiliser l'estimateur rajusté de préférence à l'estimateur non-rajusté, qui découle d'une méthode d'imputation par régression.
3. Pour la même valeur de h , le biais de TD est toujours moindre que celui de TD^* . Les écarts sont négligeables pour $h = .1$. Lorsque h augmente, TD^* se rapproche de TI plus rapidement que TD ; lorsque $h = .5$, l'écart entre TD^* et TI est négligeable sur le plan pratique.
4. La réduction du biais découlant de la substitution de TD à TI est appréciable dans tous les cas (de 55 à 82% pour le modèle A et de 67 à 92% pour le modèle B). La substitution de TD^* à TI entraîne aussi une diminution notable du biais: de 51 à 68% pour le modèle A et de 61 à 84% pour le modèle B.
5. Pour $h = .1$, TD et TD^* s'équivalent au point de vue de l'EQM quoique le second soit légèrement plus stable (c'est-à-dire, variance moindre). Pour $h = .3$ et $h = .5$, la moins grande stabilité de TD par rapport à TD^* est compensée de façon générale par un biais moindre et cette compensation est telle que TD devient préférable à TD^* au point de vue de l'EQM.

Tableau 1

Définition des estimateurs

Estimateurs		Méthode (1)		Méthode (2)	
h		TD_1		TD_1^*	
0.1		TD_3		TD_3^*	
0.3		TD_5		TD_5^*	
0.5					

Par ailleurs, nous considérons les estimateurs suivants dans l'étude de Monte Carlo:

$$TC = \bar{X} \left(\sum_s y_k / \sum_s x_k \right) \quad \text{and} \quad TI = \bar{X} \left(\sum_r y_k / \sum_r x_k \right).$$

TC est l'estimateur pour l'échantillon complet, c'est-à-dire l'estimateur par quotient selon l'hypothèse de l'absence de non-réponse, et TI est le même estimateur pour le sous-ensemble des répondants sans compensation de la non-réponse. Souignons que TI est aussi un estimateur qui découle d'une méthode d'imputation bien connue (par régression) (Cassel et coll. 1983) et qui est égal à TD lorsque h est au moins égal à l'intervalle des valeurs de x . TI n'est approximativement sans biais que si (4) est vraie. Nous verrons que le biais dépend de la différence entre les conditions définies en (4) et celles de la population à l'étude. Comme dans la simulation de la section suivante, le modèle (4) sera un "faux" modèle (c'est-à-dire que les populations étudiées sont définies par des modèles différents de (4)); par ailleurs, la simulation permet de mieux comprendre cette méthode d'imputation très simple et largement répandue.

4. ÉTUDE DE MONTE CARLO

Par une simulation de Monte Carlo, deux populations, $POP1$ et $POP2$, ont été produites selon la méthode utilisée par Särndal et Hui (1981). $POP1$ et $POP2$ comportent chacune deux strates ($S1$ et $S2$) de 500 unités chacune, et satisfont aux équations suivantes:

$$E_{\Phi}(Y_k) = \beta_1 x_{k1} + \beta_2 x_{k2},$$

$$Var_{\Phi}(Y_k) = \sigma_1^2 x_{k1} + \sigma_2^2 x_{k2}, \tag{5}$$

où $x_{k1} = x_k \partial_k$ et $x_{k2} = x_k (1 - \partial_k)$, $\partial_k = 1$ si $k \in S1$ et $\partial_k = 0$ si $k \in S2$. La différence entre les modèles (4) et (5) sert à simuler l'une des nombreuses erreurs que l'on peut commettre en définissant le modèle de superpopulation. Les caractéristiques numériques de $POP1$ et $POP2$ figurent dans le tableau 2.

Voici en bref les étapes de la simulation:

- 1) un échantillon aléatoire simple s de n ($n=50, 100$) unités est prélevé dans chaque population;
- 2) on enregistre ensuite les données de l'échantillon complet puis on crée de la non-réponse à l'aide des deux modèles paramétriques suivants:

Modèle A: $q_k = \exp(-\theta x_k)$,

Modèle B: $q_k = \theta_1^k \theta_2^{1-\partial_k}$, $\partial_k = 1$ (0) si $k \in S1$ ($S2$),

q_j par l'intermédiaire de D^* , qui est une quantité inversement proportionnelle à l'écart $|x_k - x_j|$.

La méthode (2) comporte deux difficultés; il faut en effet définir i) la fonction D^* et ii) les valeurs de ses paramètres. Dans cet article, nous définissons D^* comme une fonction de distribution d'une loi normale:

$$D^*(z) = (h^2 2\pi)^{-1/2} \exp(-z^2/2h^2); \quad z = x_k - x_j, \quad (3)$$

où l'écart-type, désigné par h , joue le même rôle que le paramètre h de l'expression (1). Lorsque h augmente dans les expressions (1) et (2), q_k tend vers la valeur constante n_p/n . Dans (1), il atteint n_p/n lorsque h est au moins égal à l'intervalle des valeurs de x . Une étude empirique visant à évaluer les propriétés des méthodes proposées a été conçue pour un très grand nombre de valeurs de h . Dans le présent article, nous ne rapportons que les résultats se rattachant à trois de ces valeurs (constantes), soit $h = 1/10$, $3/10$ et $5/10$ de l'intervalle des valeurs de x de l'échantillon. Enfin, soulignons que les équations (1) et (2) ne sont pas que des facteurs de normalisation; chacune d'elles se présente aussi comme le rapport de deux estimateurs noyaux de fonction de densité (selon l'approche de Rosenblatt (1956) pour diverses séries de valeurs de x . Comme le propose Giommi (1985b), on peut donc choisir la valeur de h en tenant compte des propositions contenues dans cette théorie.

3. MODÈLE DE SUPERPOPULATION ET ESTIMATEURS

Pour le choix de l'estimateur de Y , nous supposons un modèle de superpopulation Φ , où les valeurs de la population y_k , $k = 1, 2, \dots, N$, sont réputées former un échantillon aléatoire de telle sorte que:

$$E\Phi(Y_k) = \mu_k = \beta x_k, \\ \text{Var}\Phi(Y_k) = \sigma_k^2 = \sigma^2 x_k, \quad (4)$$

où β et Φ sont inconnues et x_k est la valeur (connue) de la variable auxiliaire X . Il est clair que ce modèle de superpopulation s'applique plus à des variables quantitatives qu'à des variables qualitatives; dans ces conditions, il conviendrait d'utiliser d'autres modèles. Par ailleurs, nous nous limitons ici à l'étude des échantillons aléatoires simples. À la condition que la variance de Y puisse être définie par l'équation en (4), Cassel et coll. (1983) ont montré que l'estimateur:

$$T = \bar{X} \left(\sum y_k / q_k \right) / \left(\sum x_k / q_k \right),$$

où \bar{X} indique la sommation par rapport à l'ensemble r et $\bar{X} = \sum_N^k x_k / N$, est approximativement non biaisé (grâce au facteur de correction q_k) même si la première équation en (4) ne définit pas réellement la relation entre X et Y . Cette situation peut être observée, par exemple, lorsque le "véritable" modèle comporte une ordonnée à l'origine ou deux coefficients de régression (voir modèle (5) ci-dessous), etc. Malheureusement, l'estimateur T ne se prête pas à des applications pratiques puisque q_k est inconnue. Nous devons donc évaluer les propriétés de cet estimateur en remplaçant q_k par son estimation calculée à l'aide de la méthode (1) ou (2). Nous allons analyser ce genre d'estimateur pour les trois valeurs de h choisies. Nous désignons les estimateurs par TD_i et TD_i^* où $i = 1, 3, 5$ (voir le tableau 1).

n unités prélevé suivant un plan d'échantillonnage $p(s)$. Pour l'estimation, nous disposons de renseignements supplémentaires représentés par les valeurs connues x_k ($k = 1, \dots, N$), d'une variable scalaire continue X (en principe, l'application des méthodes proposées pour le cas à plusieurs variables n'est pas censée poser de problèmes).

Dans l'échantillon, Y est observable uniquement pour un sous-ensemble r de n_r répondants; elle ne peut être observée pour les $n - n_r$ non-répondants. Après l'échantillonnage, l'information disponible peut être représentée de la façon suivante:

$$(k, I_k, I_{kY_k}, x_k) \quad k \in s; N, n,$$

où I_k est une variable aléatoire indicatrice telle que $E(I_k) = q_k$ et q_k est la P.R.I. Pour estimer q_k , on suppose habituellement un modèle paramétrique (Cassel et coll. 1983) de la forme:

$$q_k = q(\Theta, x_k),$$

où Θ est un paramètre (ou vecteur de paramètres) inconnu et $q(\cdot, \cdot)$ est une fonction à définir. On obtient des estimations de q_k en remplaçant Θ par son estimation $\hat{\Theta}$ dans le modèle paramétrique ci-dessus.

Dans cet article, nous estimons q_k ($k \in r$) sans utiliser de définition paramétrique de la fonction $q(\cdot, \cdot)$; nous continuons toutefois à supposer que les P.R.I. dépendent des valeurs de x_k . Deux méthodes (méthodes (1) et (2)) sont proposées.

Selon la première méthode, q_k ($k \in r$) est estimée comme taux de réponse (c'est-à-dire, la proportion de répondants) d'un groupe d'unités centré sur l'unité k et qui correspond à un intervalle approprié de valeurs de x centré sur x_k . En supposant qu'il s'agit d'un intervalle de longueur $2h_k$, q_k est estimée par le ratio suivant:

$$\hat{q}_k = \sum_{j \in r} D(x_k - x_j) / \sum_{j \in s} D(x_k - x_j), \tag{1}$$

ou

$$D(x_k - x_j) = \begin{cases} 1 & \text{si } |x_k - x_j| \leq h_k \\ 0 & \text{dans le cas contraire.} \end{cases}$$

Il est évident que l'estimation \hat{q}_k dépend de h_k ou de h si nous utilisons, comme c'est le cas ici, un intervalle fixe; dans la pratique, la détermination de la valeur de h est un problème majeur.

Selon la seconde méthode, l'estimation de q_k repose sur toutes les unités de l'échantillon et non seulement sur une partie d'entre elles. Cela élimine les inconvénients que pouvait poser une catégorisation des unités répondantes. Autrement dit, on pouvait très difficilement imaginer, pour l'estimation de q_k , que des unités aient une pondération de 1 tandis que d'autres avaient une pondération nulle. Selon la méthode (2), l'estimation est définie par:

$$\hat{q}_k = \sum_{j \in r} D^*(x_k - x_j) / \sum_{j \in s} D^*(x_k - x_j), \tag{2}$$

où D^* doit être définie. Dans ce cas, chaque valeur x_j contribue au calcul de l'estimation

Méthodes non paramétriques pour l'estimation des probabilités de réponse individuelles

ANDREA GIOMMI

RÉSUMÉ

Inspirant de l'approche de Cassel, Särndal et Wretman (1983), l'auteur aborde le problème de la non-réponse dans l'estimation de la moyenne d'une population finie. L'auteur propose tout d'abord des méthodes très simples pour estimer les probabilités de réponse individuelles; il applique ensuite ces méthodes à un modèle de superpopulation pour construire des estimateurs de la moyenne de la population. Enfin, au moyen d'une étude de Monte Carlo, il fait une première évaluation des propriétés des méthodes proposées. Les résultats de cette évaluation nous éclairent sur l'efficacité de ces méthodes.

MOTS CLÉS: Non-réponse; probabilité de réponse individuelle; méthodes non paramétriques.

1. INTRODUCTION

S'intéressant à l'estimation de la moyenne (ou total, etc.) d'une population finie en situation de non-réponse, Cassel, Särndal et Wretman (1983) ont imaginé une méthode d'estimation très générale qui repose sur la notion fondamentale de probabilité de réponse individuelle (PRI). Ils ont proposé des estimateurs qui sont déterminés en partie par un modèle de superpopulation et en partie par un modèle de réponse, c'est-à-dire un modèle qui reproduit le mécanisme de réponse et qui permet d'estimer la PRI à partir de données d'échantillon. L'estimation de la PRI est le point central de leur théorie. De fait, si le modèle de superpopulation ne convient pas parfaitement, comme cela est souvent le cas, seul un modèle de réponse approprié évitera que les estimateurs soient entachés d'un biais attribuable au plan de sondage. À l'aide d'une étude de Monte Carlo, Giommi (1985a) a montré qu'un modèle de réponse qui offrait une "bonne approximation" du "vrai" modèle de réponse pouvait éliminer virtuellement toute possibilité d'erreur systématique. Toutefois, on peut difficilement dire ce qu'est une bonne approximation et, de toute façon, le choix d'un modèle de réponse, outre qu'il est arbitraire, peut s'avérer contraignant. Une manière naturelle de contourner le problème est d'estimer la PRI par des méthodes non paramétriques. Dans cet article, nous proposons deux méthodes très simples pour estimer la PRI lorsque l'information supplémentaire dont nous disposons (et qui est censée dépendre des réponses obtenues) est représentée par une variable continue unique. Ces méthodes, qui intègrent certains éléments de la théorie d'estimation de noyaux, peuvent être considérées comme un prolongement de la méthode de compensation de la non-réponse, d'usage courant, qui consiste à répondre par des unités au moyen de cellules de correction.

Dans cet article, nous faisons une évaluation empirique de ces méthodes et analysons les biais et l'efficacité des estimateurs correspondants.

2. ESTIMATION DES PROBABILITÉS DE RÉPONSE INDIVIDUELLES

Considérons une population de N unités identifiées k ($k = 1, 2, \dots, N$), et une variable Y étudiée, dont nous voulons estimer la moyenne $Y = \sum_k y_k / N$ à l'aide d'un échantillon s de

Les quatre autres articles du présent numéro traitent de l'élaboration et de l'application de méthodes et de procédures relatives aux probabilités de réponse dans le cadre d'une enquête, aux critères d'arrondissement visant à protéger la confidentialité, à la collecte et l'analyse des données aux fins d'enquêtes rétrospectives et à l'estimation de la variance dans l'enquête sur la population active du Canada.

Toute enquête donne lieu à des problèmes de non-réponse. On les résoud ordinairement par imputation ou ajustement, en supposant que la non-réponse est aléatoire à l'intérieur des catégories d'imputation ou d'ajustement. Les estimations ainsi produites sont généralement biaisées si l'hypothèse précitée n'est pas satisfaite. On a proposé diverses méthodes d'estimation des probabilités de réponse au moyen de modèles, notamment la méthode de Cassel, Sarnadal et Wretiman (CSW), mais ces méthodes ne sont pas efficaces lorsque le modèle supposé est inadéquat. Dans "Méthodes non paramétriques pour l'estimation des probabilités de réponse individuelles", Giommi décrit des méthodes non paramétriques d'estimation des probabilités de réponse avec utilisation de l'information auxiliaire, fournissant ainsi une alternative à l'estimateur de CSW qui résiste bien aux problèmes causés par des modèles de population et de réponse qui ne sont pas valables. Les estimateurs obtenus donnent de bons résultats dans les simulations de Monte Carlo.

L'arrondissement aléatoire est utilisé pour garantir le caractère confidentiel des renseignements sur des personnes dans les agrégats statistiques. Dans le contexte du recensement du Canada de 1971, Nargundkar et Saveland ont élaboré un processus d'arrondissement sans biais dans lequel la valeur attendue des données arrondies est la même que celle des données non arrondies. Fellagi (TF, 1975) a proposé l'arrondissement aléatoire contrôlé qui, en plus d'être sans biais, conserve aux données leur caractère additif. Plusieurs autres articles sur la question ont été publiés depuis, dont celui très récent de Cox (JASA, 1987), qui généralisent et étendent les applications à d'autres domaines. Dans "Estimations fondées sur des données arrondies aléatoirement", Withers donne une expression de la variance des estimations sans biais des probabilités de cellule et présente une comparaison de l'efficacité faisant intervenir les méthodes d'arrondissement utilisées en Australie, au Royaume-Uni, en Nouvelle-Zélande et au Canada. Il étend aussi ses résultats à toute fonction lisse des probabilités de cellule, à appliquer à divers secteurs de la statistique.

Dans "Estimation de la variance pour l'enquête sur la population active du Canada", Choudry et Lee décrivent les études effectuées en vue de choisir un estimateur de la variance des estimations itératives par le quotient provenant de l'enquête sur la population active. Ils donnent les résultats d'une comparaison de trois estimateurs de la variance pour le plan d'échantillonnage par groupe aléatoire (Keyfitz, Rao-Hartley-Cochran et Rao). Même si elle est inférieure aux deux autres méthodes pour ce qui est du biais et de la stabilité, la méthode de Keyfitz est recommandée parce qu'elle est simple à utiliser.

Dans "La fiche AGEVEN: un outil pour la collecte des données rétrospectives", Antioine, Bry et Diouf décrivent les techniques utilisées pour recueillir des données sur la natalité et la mortalité auprès des femmes de Pikine, une banlieue de Dakar, au Sénégal. La démarche rétrospective suivie comporte le placement d'événements observés (principalement les naissances et les décès) dans leur contexte socio-économique; elle permet, selon les auteurs, "de mieux évaluer la relation entre l'insertion urbaine et les changements de comportement démographique". On constate aisément en analysant les données de l'enquête que le taux de mortalité est plus élevé chez les enfants nés dans des villages ruraux que chez ceux qui naissent à Pikine. Il est bien connu que la stratégie de Hansen et Hurwitz est inférieure à celle de Horvitz et Thompson dans le cas d'un certain nombre de procédures d'échantillonnage avec PSPT (probabilité de sélection proportionnelle à la taille). Dans le dernier article du numéro, dans la section "Communications brèves", Prabhu-Ajgaonkar donne de ces résultats des démonstrations beaucoup plus simples que celles qu'on trouve actuellement dans la littérature spécialisée.

Dans ce numéro

Deux nouvelles rubriques paraissent dans la présente livraison de *Techniques d'enquête*. **Dans ce numéro** résume les articles que renferme le numéro, et elle paraîtra régulièrement. La deuxième s'intitule **"Communications brèves"** et elle sera publiée à l'occasion. Vous trouverez dans ce numéro neuf articles dont quatre consacrés à des **méthodes d'estimation et de pondération**; deux de ces derniers traitent des estimations sur les familles. C'est grâce à l'initiative de Fritz Scheuren et à son aide lors de la rédaction que cette section spéciale paraît dans la présente livraison de *Techniques d'enquête*.

Les trois premiers articles de la section spéciale traitent (au moins partiellement) des méthodes des moindres carrés pour pondérer des données d'enquête. Il y a là une certaine ironie du sort: dans leur article de 1940, Deming et Stephan ont présenté l'ajustement proportionnel itératif comme une méthode rapide et pratique pour obtenir des approximations des estimations obtenues en minimisant une fonction quadratique des cellules d'un tableau de contingence, sous réserve de restrictions marginales. Cette technique s'est passablement répandue comme mode de pondération des données d'enquête, on l'appelle la "méthode itérative du quotient". Dans "Méthode alternative pour ajuster les estimations de la Current Population Survey aux chiffres de population", Copeland, Pelzmeier et Hoy comparent un estimateur fondé sur la méthode itérative du quotient et un estimateur fondé sur les moindres carrés généralisés, sous réserve dans les deux cas des mêmes restrictions marginales. Ils comparent des estimations de caractéristiques individuelles obtenues lors de la Current Population Survey, une enquête sur les ménages exécuté par le Bureau of the Census des États-Unis. Ils notent que les estimations produites par les deux méthodes sont semblables.

La plupart des méthodes actuelles de pondération des données d'enquête sur les ménages donnent des poids différents pour les divers membres d'un ménage donné. Un seul poids par ménage, en plus d'être conceptuellement attrayant, éliminerait les différences répétées souvent embarrassantes entre les estimations par personne et par famille. Alexander, dans "Une classe de méthodes utilisant des chiffres de population dans la pondération des ménages", considère une catégorie de méthodes fondées sur la notion de "distance minimum conditionnelle" (dont celle des moindres carrés généralisés) qui en fait produisent un seul poids par famille tout en respectant les chiffres de population pour les individus. Les propriétés de ces méthodes en cas de sous-dénombrement sont ensuite étudiées au moyen de quelques modèles de couverture simples.

Lemaître et Dufour, dans "Une méthode intégrée de pondération des personnes et des familles", proposent un estimateur par la régression qui produit également un seul poids par ménage et qui, sous certaines réserves générales, équivaut à l'estimateur par les moindres carrés généralisés. En se servant de données de l'enquête sur la population active du Canada, ils atteignent un degré d'efficacité beaucoup plus élevé dans leurs estimations relatives aux familles et un peu plus élevé dans leurs estimations de personnes, comparativement aux méthodes actuelles. Dans le dernier article de la section spéciale, intitulé "Variante de la méthode itérative du quotient", Oh et Scheuren décrivent une méthode d'estimation qui ressemble à la méthode itérative usuelle. Leur méthode peut être utilisée lorsque les totaux de la population sont disponibles non seulement pour les marges mais aussi pour les cellules intérieures d'un tableau multidimensionnel. Elle combine l'estimation conventionnelle par le quotient pour les cellules contenant les échantillons de grande taille et l'estimation itérative par le quotient pour les cellules dans lesquelles l'échantillon est petit (ou inexistant). Dans une application concernant un échantillonage de déclarations d'impôt de sociétés, la méthode Oh-Scheuren a produit des estimations plus efficaces que la méthode itérative du quotient conventionnelle. Les auteurs soulignent le fait qu'il reste encore du travail à faire, notamment comparer leur méthode aux méthodes conventionnelles de regrouper les données, avant de songer à répandre cette approche.

TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada
 Volume 13, numéro 2, décembre 1987

TABLE DES MATIÈRES

Dans ce numéro	135
A. GIOMMI	
Méthodes non paramétriques pour l'estimation des probabilités de réponse individuelles	137
C.S. WITHERS	
Estimations fondées sur des données arrondies aléatoirement	145
G.H. CHOUDHRY et H. LEE	
Estimation de la variance pour l'enquête sur la population active du Canada	157
P. ANTOINE, X. BRY, et P.D. DIOUF	
La fiche «AGEVEN»: un outil pour la collecte des données rétrospectives	173
Section Spéciale - Méthodes d'estimation et de pondération	
K.R. COPELAND, F.K. PEITZMEIER, et C.E. HOY	
Méthode alternative pour ajuster les estimations de la Current Population Survey aux chiffres de population	183
C.H. ALEXANDER	
Une classe de méthodes utilisant des chiffres de population dans la pondération des ménages	193
G. LEMAÎTRE et J. DUFOUR	
Une méthode intégrée de pondération des personnes et des familles	211
H.T. OH et F. SCHEUREN	
Variante de la méthode itérative du quotient	221
Communications brèves	
S.G. PRABHU-AJGAONKAR	
Comparaison de la méthode de Horvitz-Thompson et de la méthode de Hansen-Hurwitz	233
Remerciements	237

TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada

La revue Techniques d'enquête est répertoriée dans The Survey Statistician et Statistical Theory and Methods Abstracts. On peut en trouver les références dans Current Index to Statistics

COMITÉ DE DIRECTION

Président

G.J. Brackstone

Membres

B.N. Chinnappa

G.J.C. Hole

C. Patrick

F. Mayda (Directeur de la production)

COMITÉ DE RÉDACTION

Rédacteur en chef

M.P. Singh, *Statistique Canada*

Rédacteurs associés

K.G. Basavarajappa, *Statistique Canada*

D.R. Bellhouse, *University of Western*

Ontario

L. Biggert, *Université de Florence*

D. Binder, *Statistique Canada*

E.B. Dagum, *Statistique Canada*

W.A. Fuller, *Iowa State University*

J.F. Gentleman, *Statistique Canada*

D. Holt, *University of Southampton*

Rédacteurs adjoints

J. Armstrong, J. Gambino et H. Lee, *Statistique Canada*

POLITIQUE DE RÉDACTION

La revue Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception de sondages, les erreurs dans les enquêtes, l'évaluation de différentes sources de données et de méthodes de collecte, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration de données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

La revue Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à en faire parvenir le texte au rédacteur en chef, M. M.P. Singh, Division des méthodes d'enquêtes sociale, Statistique Canada, 4^e étage, Edifice Jean-Talon, Tunney's Pasture, Ottawa (Ontario), Canada K1A 0T6. Prière d'envoyer deux exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

Abonnement

Le prix de la revue Techniques d'enquête (catalogue n° 12-001) est de 20,00\$ par année au Canada, et de 23,00\$ par année à l'étranger (paiement en dollars canadiens ou l'équivalent). Prière de faire parvenir votre demande d'abonnement à: Section des ventes des publications, Statistique Canada, Ottawa (Ontario), Canada K1A 0T6. Un prix réduit, soit 10,00\$ (E.-U.) (\$14,00 Can.) est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticien et d'Enquête et la Société Statistique du Canada. Veuillez envoyer votre demande d'abonnement directement à l'organisation.

TECHNIQUES D'ENQUÊTE

UNE REVUE DE STATISTIQUE CANADA
DÉCEMBRE 1987

Publication autorisée par
le ministre des Approvisionnements
et Services Canada
©Ministre des Approvisionnements
et Services Canada 1988

Le lecteur peut reproduire sans autorisation des
extraits de cette publication à des fins d'utilisation
personnelle à condition d'indiquer la source en
entier. Toutefois, la reproduction de cette publication
en tout ou en partie à des fins commerciales ou de
redistribution nécessite l'obtention au préalable
d'une autorisation écrite des Services d'édition,
Agent de droit d'auteur, Centre d'édition du gouvernement
du Canada, Ottawa, Canada K1A 0S9.

Mai 1988

Prix: Canada, \$20.00 par année
Autres pays, \$23.00 par année

Païement en dollars canadiens ou l'équivalent
Catalogue 12-001, vol. 13, n° 2

ISSN 0714-0045

Ottawa

Canada

VOLUME 13, NUMÉRO 2
DÉCEMBRE 1987

UNE REVUE
DE
STATISTIQUE CANADA

TECHNIQUES D'ENQUÊTE





Statistics Canada Statistique Canada

SURVEY METHODOLOGY

A JOURNAL
OF
STATISTICS CANADA



VOLUME 14, NUMBER 1
JUNE 1988

Canada

SURVEY METHODOLOGY

A JOURNAL OF STATISTICS CANADA

JUNE 1988

Published under the authority of
the Minister of Supply and
Services Canada

©Minister of Supply
and Services Canada 1988

Extracts from this publication may be reproduced
for individual use without permission provided the
source is fully acknowledged. However, reproduction
of this publication in whole or in part for purposes
of resale or redistribution requires written permission
from the Publishing Services Group, Permissions
Officer, Canadian Government Publishing Centre,
Ottawa, Canada K1A 0S9

September 1988

Price: Canada, \$20.00 a year
Other Countries, \$23.00 a year

Payment to be made in Canadian funds or equivalent

Catalogue 12-001, Vol. 14, No. 1

ISSN 0714-0045

Ottawa

SURVEY METHODOLOGY

A Journal of Statistics Canada

The Survey Methodology Journal is abstracted in The Survey Statistician and Statistical Theory and Methods Abstracts and is referenced in the Current Index to Statistics.

MANAGEMENT BOARD

Chairman	G.J. Brackstone	
Members	B.N. Chinnappa	R. Platek
	G.J.C. Hole	D. Roy
	C. Patrick	M.P. Singh
	F. Mayda (Production Manager)	

EDITORIAL BOARD

Editor M.P. Singh, *Statistics Canada*

Associate Editors

K.G. Basavarajappa, <i>Statistics Canada</i>	G. Kalton, <i>University of Michigan</i>
D.R. Bellhouse, <i>U. of Western Ontario</i>	M.N. Murthy, <i>Applied Statistics Centre, India</i>
L. Biggeri, <i>University of Florence</i>	W.M. Podehl, <i>Statistics Canada</i>
D. Binder, <i>Statistics Canada</i>	J.N.K. Rao, <i>Carleton University</i>
E.B. Dagum, <i>Statistics Canada</i>	D.B. Rubin, <i>Harvard University</i>
W.A. Fuller, <i>Iowa State University</i>	I. Sande, <i>Statistics Canada</i>
J.F. Gentleman, <i>Statistics Canada</i>	C.E. Särndal, <i>University of Montreal</i>
M. Gonzalez, <i>U.S. Office of Management and Budget</i>	F.J. Scheuren, <i>U.S. Internal Revenue Service</i>
D. Holt, <i>University of Southampton</i>	V. Tremblay, <i>Statplus, Montreal</i>
	K.M. Wolter, <i>U.S. Bureau of the Census</i>

Assistant Editors

J. Armstrong, J. Gambino and J.-L. Tambay, *Statistics Canada*

EDITORIAL POLICY

The Survey Methodology Journal publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

The Survey Methodology Journal is published twice a year. Authors are invited to submit their manuscripts in either of the two official languages, English or French to the Editor, Dr. M.P. Singh, Social Survey Methods Division, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Two nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of the Survey Methodology Journal (Catalogue No. 12-001) is \$20.00 per year in Canada, \$23.00 per year for other countries (payment to be made in Canadian funds or equivalent). Subscription order should be sent to: Publication Sales, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6. A reduced price of US \$10.00 (\$14.00 Can.) is available to members of the American Statistical Association, the International Association of Survey Statisticians, and the Statistical Society of Canada. Please subscribe through your organization.

SURVEY METHODOLOGY

A Journal of Statistics Canada
Volume 14, Number 1, June 1988

CONTENTS

In This Issue	1
J. RYTEN	
Errors in Foreign Trade Statistics	3
L. KISH	
Multipurpose Sample Designs	19
P. LAVALLÉE and M.A. HIDIROGLOU	
On the Stratification of Skewed Populations	33
D. KASPRZYK	
Research Issues in the Survey of Income and Program Participation	45
D. SCHNELL, W.J. KENNEDY, G. SULLIVAN, H.J. PARK, and W.A. FULLER	
Personal Computer Variance Software for Complex Surveys	59
Special Section – Census Coverage Error	
G. DIFFENDAL	
The 1986 Test of Adjustment Related Operations in Central Los Angeles County	71
N. SCHENKER	
Handling Missing Data in Coverage Estimation, with Application to the 1986 Test of Adjustment Related Operations	87
H. HOGAN and K. WOLTER	
Measuring Accuracy in a Post-Enumeration Survey	99
P.P. BIEMER	
Modeling Matching Error and its Effect on Estimates of Census Coverage Error ..	117

In This Issue

Four of the nine papers in this issue deal with **Census Coverage Error**. These papers and others that will appear in the December 1988 issue of the Journal are valuable additions to the rapidly growing literature on this topic. Kirk Wolter's initiative was very helpful in arranging for these special sections.

Census counts are known to be inaccurate due to coverage error and this problem has recently attracted a great deal of attention among both policy makers and statisticians – academics and practitioners alike. Consequently, methods of measuring the quality of census counts including the limitations of such methods, adjustment techniques (both design and model based) to improve the quality of population figures, the impact of an undercount on various government programs and other related studies have assumed increasing importance. In many countries, evaluation studies to measure coverage are carried out during or following each census. In Canada, for example, the Reverse Record Check is the most important study undertaken to measure census undercount. Similarly, in the United States since the 1950 Census, a Post-Enumeration Survey (PES) has been one of the important vehicles used to evaluate census coverage.

In 1986, the U.S. Bureau of the Census carried out a study called Test of Adjustment Related Operations (TARO) in Los Angeles to test a new PES design. Three papers in the special section – those of Diffendal, Schenker, and Hogan and Wolter – thoroughly evaluate the methods and procedures used in this new PES, and provide an in-depth analysis of research findings, as well as the issues and achievements of the TARO. Diffendal presents an overview of the test, describing its methodological and operational aspects. His paper also contains a brief historical description of coverage measurement studies in the United States and recent events leading to the elaborate studies by the U.S. Bureau of the Census.

Schenker discusses three methods for dealing with missing data: hot deck imputation, logistic regression modeling and weight adjustment. The choice of method depends on the type of missing data. For example, logistic regression is used to impute values for binary characteristics. Using TARO data, the author compares coverage error estimates obtained under different imputation models.

Hogan and Wolter present a detailed discussion of the potential sources of error in the new PES estimates and assess the impact of individual error components as well as the overall impact of errors on TARO data. Based on their findings the authors conclude that, in practice, the PES estimates may be "more accurate than original census estimates for some areas, with equal or nearly equal accuracy for most other areas".

The fourth paper in the special section, Biemer's "Modeling Matching Error and Its Effect on Estimates of Census Coverage Error" deals with the specific problem of PES-Census matching. The author considers three increasingly complex models and examines the impact of matching on the PES estimates. Implications of the findings for the 1990 Census are discussed.

The other five papers in this issue deal with errors in foreign trade statistics, design issues in multipurpose surveys, stratification of skewed populations, the Survey of Income and Program Participation conducted by the U.S. Bureau of the Census and personal computer software for variance estimation in complex surveys.

In "Errors in Foreign Trade Statistics" Ryten discusses the sources of errors in foreign trade statistics as well as procedures for reducing these errors. He proposes the reporting of the levels of uncertainty in detailed figures. The author explains the causes of discrepancies in counterpart

trade statistics and analyses their relative importance. Based on the results of a study of the import and export data from a World Trade database created at Statistics Canada, the author raises serious questions about the comparability of counterpart data at detailed levels of commodity classification. A program to improve the quality of foreign trade statistics is proposed and arguments are made for providing users with more factual information about data quality.

In practice multipurpose uses are often made of data obtained from most surveys. However, research literature and text books usually avoid the discussion of "multipurpose sample designs". This important topic is addressed by Kish in his paper. He first presents a hierarchy of purposes and then discusses various conflicting requirements in designing a multipurpose survey. Ten areas of conflict, including determination of sample size and its allocation to domains and strata, bias to sampling error relationship, choice of stratification variables and continuity of data over time are examined. Solutions are proposed for each area and the use of compromise designs rather than designs that are optimal for a single purpose is stressed. Some proposals are less rigorous and are presented to stimulate further research on this topic.

An iterative algorithm for the stratification of skewed populations under power allocation (an allocation proportional to the stratum total raised to a low-valued positive power) is given by Lavallée and Hidioglou in their paper "On the Stratification of Skewed Populations". An empirical study is presented, comparing the suggested allocation with other allocation methods using data from the Annual Retail Trade and Wholesale Trade Surveys conducted by Statistics Canada.

The Survey of Income and Program Participation (SIPP) is an ongoing household survey conducted by the U.S. Bureau of the Census. In "Research Issues in the Survey of Income and Program Participation", Kasprzyk reviews methodological and statistical issues related to the SIPP. The paper examines four topics of special interest related to panel surveys of families and individuals. These are questionnaire design, data collection, response error and sampling and estimation issues for longitudinal concepts. The paper describes the important issues, provides references to studies conducted to address those issues and summarizes the main results of the studies.

In the paper "Personal Computer Variance Software for Complex Surveys", Schnell, Kennedy, Sullivan, Park and Fuller describe a program called PC CARP, developed to analyse data from complex surveys. This program has found applications, in particular, in many developing countries. The features and capabilities of the system are briefly described.

The Editor

Errors in Foreign Trade Statistics

JACOB RYTEN¹

ABSTRACT

In spite of the comparative ease with which studies of error in foreign trade statistics could be conducted, there are few attempts to quantify their size, origin, distribution, and change over time. Policy makers and trade negotiators have little notion of how uncertain these statistics are in spite of their great detail. This paper takes advantage of a World Trade Database developed by Statistics Canada to examine and quantify discrepancies in existing foreign trade statistics.

KEY WORDS: Foreign trade; Bilateral trade balances; Errors.

1. INTRODUCTION

This paper discusses some of the underlying causes of errors in foreign trade statistics; difficulties in detecting errors; ways of conveying the uncertainty in the detailed figures; and a proposal to improve the quality of the data.

There has not been much written about error in foreign trade statistics since Allen and Ely (1953) co-edited a book on these statistics thirty five years ago. Some attention has been paid to accounting matters — inclusions and exclusions, demarcation of boundaries, valuation, etc. (United Nations, 1982) — and most of all to classification. In fact, one of the biggest changes in trade classification ever has just been introduced (United Nations 1986) in order to make foreign trade data more comparable among countries. But perhaps because these statistics rely on a complete accounting of all merchandise transactions that take place across borders in any period of time and this accounting is enforced by a policing agency — customs administration — there is a widespread belief that there is not much measurable error left. The lack of analysis of error in these statistics supports this contention.

Periodically, it has come to the attention, particularly of statistical offices in international agencies, that there is a serious error in the reporting of trade between pairs of countries. At its eighteenth session, the United Nations Statistical Commission (1974) was formally informed of the reconciliation of trade statistics between the United States and Canada. This followed the detection of some embarrassing differences in the bilateral trade balance between the two countries. Thereafter, and at various times, issues involving Singapore and Malaysia, Singapore and Indonesia, and any of a number of non-EEC countries and the Netherlands were brought up for discussion at international agencies that were more specifically interested in trade matters. Moreover, countries which felt that they were losing control over the quality of their foreign trade statistics — typically third world countries — have attempted to piece back their own numbers by reference to those of their principal trading partners. But there is no evidence that any of these expressions of concern has ever resulted in a systematic programme to detect, measure and reduce error in the underlying statistics.

There are few obvious alternative explanations for this lack of action other than the belief that there is no error. Foreign trade statistics are among the very few where there can be a comparison of two measurements of the same transaction derived in virtually the same detail

¹ Jacob Ryten, Assistant Chief Statistician, Statistics Canada, 13-B8 Jean Talon Building, Tunney's Pasture, Ottawa, Ontario K1A 0T6.

using the same procedures, by two independent record takers. The differences that result when these comparisons are made have been referred to in the literature going back to almost the first world war (Coats 1926). And yet, they have not resulted in proposals to incorporate the results of these comparisons in any report on the quality of the underlying statistics. One of the deterrents to pursue these comparisons systematically may have been the volume of computing they entail and the expense involved. Another may be the depth of knowledge that is required of counterpart statistical systems which, in addition to being described in some instances in a foreign language, usually involve very specific administrative and legal provisions which are not comparable from country to country.

The deterrents to systematic comparisons have changed somewhat in Statistics Canada where a world trade data base has been established. Its contents are detailed trade statistics of the countries that report data in machine-readable form to the United Nations Statistical Office (UNSO). UN member countries undertake, under the terms of membership, to report a number of key statistics to the UNSO in the manner specified by the UN Secretary General. These statistics include foreign trade statistics broken down by country and commodity, with the latter in either the full detail of the Standard International Trade Classification (SITC) or its equivalent Customs Cooperation Council Nomenclature (CCCN). Annual reports in machine-readable form go back to the early sixties.

The world trade data base was created to support Canadian negotiators involved in the current round of multilateral tariff reductions and also to help Canadian exporters and importers get a better understanding of the markets and suppliers with which they deal. Its shortcomings are that it is not complete. The centrally planned economies either fail to report or else only provide very aggregate data; many of the third world countries experience serious delays in processing their Customs records as a result of which there is still much missing in recent years; not all countries report on the same vintage of the SITC; and there is a fair amount of variation in the concepts and definitions adopted by different countries.

But these shortcomings are more than offset by the fact that the computing involved in comparing trade statistics is now manageable; that a very large proportion of world trade only involves the western countries and is reported currently; and that the latter have moved to progressively more comparable conceptual frameworks. Taking these elements into account, a world trade data base can be used to display the results of comparing counterpart trade statistics and this in turn should help statistical agencies to become more conscious of the strengths and weaknesses of their merchandise imports and exports data. This is a necessary condition to improve the reliability of trade statistics. Given the attention that is currently paid to these data, statistical agencies throughout the world are well advised to make the improvements suggested by bilateral comparisons of counterpart data even if they can only do so gradually.

In the next sections, there is a review of the principal causes of discrepancies in counterpart statistics and of what steps can be taken to estimate their relative importance in particular situations.

2. TRADE TRANSACTION RECORDS: ERRORS AND DIFFERENCES IN COUNTERPART RECORDS

Underlying two counterpart trade records, there is, in most cases, one single documented transaction. An exporter has made a sale and invoiced the purchaser accordingly. That invoice is likely to contain the essential facts about the transaction which includes a description of the product(s) sold, the corresponding value and quantity, the terms and conditions of the sale, an identification of the purchaser and of the purchaser's residence and a date on which the transaction took (or will take) place. This record generates a number of related records, some

derived by transforming the basic information in some prescribed manner and others through record linkage with related records. Examples of the latter include a description of how the products transacted were moved from the place of sale to the place of purchase and how much that cost, the cost of insuring the shipment, what amounts were charged to the two parties to the transaction because of duties, sales taxes, consular charges etc.; and of course, the form and date in which the purchase was settled.

The transformations of the basic information have to do with conventions regarding the way in which this basic information is recorded and the documentation of the different stages of the transaction over time. These transformations are not standard across countries. The conventions that rule them are either embodied in Customs law or else in the administrative regulations that govern Customs record keeping. They give rise to the documents that form the basis of foreign trade statistics. One set of documents is kept by the country of sale; and the other by the country of purchase. In practice these documents differ in spite of relating to what is in principle and in fact the same commercial transaction.

Firstly, they differ in time. Even between adjacent countries or in cases where air transport is involved, differences in time are not trivial. They arise because the chain of links that make up the transaction is long — bringing the shipment to the point from which the international carrier will depart; warehousing while waiting for international transport; arriving at the point of destination; warehousing while waiting to clear Customs formalities; and while this is going on, filing documents at different stages and having them recorded on the basis of different conventions. Also, in one country the time of transaction may be recorded as the time the invoice is received in the importing country and in another as the time amounts owing to the Customs administration are paid.

Secondly, in one country the recording of the value of the purchase may include all costs of international transportation and insurance; whereas in another these may be kept separately. Thirdly, in one country the transaction may be imputed not to the country from which the invoice was issued but rather to the country where the product was grown, extracted or manufactured; whereas in another, it is the residence of the seller that decides the country assignment. Political stances can also affect the way a country is identified on the records. Fourthly, customs regulations can bias the way imports or exports are recorded. Fifthly, there are data coding and processing errors. And finally, the units in which the quantities are reported can cause inconsistencies. The following sections provide additional detail on these factors.

i) Differences between exports and imports records: timing

Customs administrations will normally file records in a variety of ways: by country of origin; by the identification of the importing business or its agent; and by time of receipt. But there are at least four key events involved in an import transaction all of which may be recorded but only one of which will be chosen as the date for retrieval and statistics. The choice of the date is not subject to statistical standardization but rather to how customs views its prime function and to the technical capacity to store alternatives. Clearly, if one country chooses as its date to record exports the time when the forwarding agency prepares an export document; and the counterpart country chooses as time for imports the date when all duties and other dues are settled, the possible lag between the recording of exports and the corresponding imports is a maximum.

ii) Differences between exports and imports records: values

Value differences have long stood in the way of systematic comparisons so it is best to review them and assess their relative importance. The valuation of the transaction that is to say, the price at which it is recorded for purposes of customs administration — is critical. Many countries (most?) record the value of an import including the cost of international transport and

insurance relating to the shipment. Most countries record the value of the counterpart exports excluding these components. There are additional variations: some countries include portions of inland transport and insurance and some countries exclude harbour costs from costs of international transport. But these differences only present a marginal increase in the difficulty of comparing counterpart records. Transactions involving related commercial partners as in the case of multinational enterprises trading internationally pose a problem of valuation which is solved in different ways in different countries. It is possible that this source of difference will outstrip all others in the years to come.

iii) *Differences between exports and imports records: country*

There is the matter of country crediting which can introduce some of the more puzzling differences in any systematic programme of comparisons. As an exporter, a country can count as an export any sale of goods that has to cross its customs boundaries to reach its point of destination, independently of whether it was substantially changed or is being sold in the exact same form in which it was purchased from some other country. However, as an importer a country may decide to impute a purchase to the country where the last substantial transformation (normally "substantial" has a precise definition in law) took place. Accordingly in the case of three hypothetical countries, A, B, and C where A has exported some goods to B and B has exported the same goods (perhaps transformed) to C, the statistics may be recorded in any of many possible ways with different consequences, as shown in the table below.

The symbols "*x*" and "*m*" denote respectively value of exports to and imports from the partner country (second upper case letter) as recorded by the reporting country (first upper case letter).

Accordingly,

A_xB = Value of exports from A to B as recorded by A

A_mB = Value of imports from B to A as recorded by A

	Recorded as exports		Recorded as imports	Consequence
i)	$A_xB + B_xC$	—	$B_mA + C_mB$	Consistent and complete
ii)	$A_xB + B_xC$	—	$B_mA + C_mA$	Overcrediting of A by importers
iii)	$A_xC + B_xC$	—	$B_mA + C_mB$	Overcrediting of C by exporters
iv)	A_xC	—	C_mA	Consistent but incomplete
v)	A_xC	—	C_mB	No crediting of A by importers
vi)	A_xB	—	C_mA	No crediting of C by exporters

The different cases indicate that some reporting countries credit their exports to the first and others to the last known destination; that some importing countries credit imports to the country of origin and others to country of consignment; and some exporting countries count as exports whatever leaves their national territory irrespective of the degree of transformation

to which the goods may be subject. The differences involved in these approaches are not trivial matters in days of free trade agreements, Customs unions, free trade zones and other arrangements to stimulate transborder trade. For each of these arrangements, a separate statistical convention is needed to accommodate the effect of the agreement on customs record keeping. Crediting partner countries in inconsistent ways is only one source of discrepancy in bilateral or multilateral comparisons. The other is due to inconsistent geographic classification.

In fact, many countries embody their stance in international politics in their standard geographical classifications. Accordingly, there are differences that arise from inconsistent geographic definitions of partner countries. Most Latin American countries treat Puerto Rico as a separate origin or destination from the United States. Virtually each OECD member country has a different treatment of partner countries in Africa. Some lump them together by their colonial origins and others by geographic neighbourhood. Similar inconsistencies arise in the treatment of the Caribbean and South Pacific islands. The Economic Union of South Africa is treated in the statistics in ways which often reflect the reporting country's view of an embargo on commercial ties with South Africa itself. Moreover, not all countries track the changes in the political status of their trading partners with the same zeal so that not all catch up with newly created independent nations as quickly as desirable in order to conduct statistical comparisons.

iv) *Differences between exports and imports records: Customs administration*

There is another important difference that arises because the attention paid to exports by Customs administrations is less than what their mandate requires they pay to imports. The reporting of individual exports shipments may be consolidated in the interests of paper burden and brought into line with the manifests or other transport documents handled by the carrier. In the case of imports, the objective is to get reporting in sufficient detail to allow Customs to apply the right duties and other taxes. One consequence is that in the case of exports, low value components of a mixed shipment are more likely to be classified under the same heading as the major component whereas in the case of imports the chances are that they will be classified independently.

This difference in interest that can be ascribed to the mandate of a Customs administration has other substantial effects on the quality of exports and imports documents. On the one hand there is evidence that the extent of underreporting of exports which affected United States overland exports to Canada is not confined to North America. Almost twenty years ago the United Kingdom launched a massive programme that consisted in matching shipping manifests to export documents because of a perceived rate of underreporting of some one to two per cent of the total. On the other hand, there is a presumption that the description of exported products is unbiased (unless it covers up illegal shipments) whereas the descriptions of imported goods may be biased because they aim at minimizing the rates of duty for which the imports are liable.

In addition to these sources of difference, which are due to the different legal and administrative transformations to which the original record is subject, there are others which are more variable and more selective in terms of the records to which they apply. Examples are the treatment of low value shipments (they are defined as below different thresholds and are excluded, included, or sampled with varying rates) and the treatment of commodities that have important service elements such as recorded audio and video tapes, architects' blueprints; computing software recorded on magnetic tape; repairs and maintenance etc.

v) *Differences between exports and imports records: coding and data processing*

Virtually all classes of information that are included in the basic records kept by Customs reflect the application of a classification or a code to an actual situation. The way to ensure

consistency of coding is by ruling on borderline cases and ensuring that the accumulated rulings form something akin to case law — a body of decisions to be made accessible to coders and by which they should be governed. But the only central dispenser of rulings is the Secretariat of the Customs Cooperation Council in Brussels and it can neither be consulted by member countries on a day to day basis nor can its decisions go beyond a certain level of generality. For this reason, there are systematic differences in interpreting and applying standard codes sometimes within the same country, let alone among different countries.

In addition, there are inconsistencies due to errors at the data processing stage and as a consequence of the systems put in place to reduce their impact. For example, there are errors in interpreting Customs legislation and in coding source information that creep in at the stage when importers or exporters inform their authorities of an impending shipment; errors at the stage of data capture; and errors of coding within the statistical agency. The standard protection against these errors is the institution of review and editing systems that rely to differing extents on clerical inspection and review and on computerized detection and imputation. Although it is very likely that there are other sources for inconsistency, the issues reviewed above are the most frequently cited ever since these matters were first described in the literature (Coats 1926), and probably are the most important explanations of the differences in counterpart figures.

vi) *Differences between exports and imports records: quantities, a special variable*

Unlike values, reported quantities are not affected by the inclusion of transport costs nor are they biased in order to minimize tax liabilities (although if values are miscoded to lower duty categories they will drag the matching quantities along). Unfortunately, there are other problems associated with the recording and use of quantities that greatly reduce the value of these statistics for error detection. For example, quantities can apply to either an entire shipment in which case they are usually expressed as a gross weight or else to a specific commodity in which case they are expressed as either net weights or in any other appropriate unit (length, surface, volume) including, in the case of complex commodities, numbers of units.

While quantity measurements in gross or net weights are comparable across countries, their use is limited by the heterogeneity of the shipments to which they refer. Quantities expressed in other units are limited by the the variety of units used and, more importantly, by the fact that they cannot be aggregated in the commodity classification and the levels to which they apply are much too detailed for inter-country comparison, given our current state of knowledge. Nonetheless, there is a use for these units in matching trade in raw materials particularly if in conjunction with values, they are used to track changes in unit values. In fact, a proposal for an international study of errors in trade relying chiefly on the matching of unit values was made to the eighteenth session of the United Nations Statistical Commission (1974). However, member countries did not feel the possible benefit justified the expected cost. At this stage, Statistics Canada's world trade data base does not include quantity information so that the applications of quantity statistics have not yet been studied.

3. A PROGRAMME TO MEASURE ERRORS

The causes of errors have been known for many years (Coats 1926). A proper attempt at quantification was made in the first reconciliation project between the United States and Canada in the early 70's. But to this day that is a very large proportion of what is known about errors in the foreign trade statistics and obviously suffers from the fact that it concerns trade between two adjacent countries and only those two countries. Given the fact that international data

bases such as the one that Canada has will likely become more popular and that they will be provided with a variety of analytical software, it is timely to speculate on what might be done to improve trade statistics, or failing improvement, at least to inform users about the limitations of foreign trade data. It is not likely that at this stage, with the descriptive information that is currently available, users in any country realize by just how much the long term trends in trade statistics might be off, or how the monthly movements in their national trade balances are affected, and most important, how prone to error is information at detailed commodity level.

Clearly, the flow $A_x B$ should be the same as $B_m A$ so long as all shipments and their recording is instantaneous, the basis of valuation is the same for the two partners for the same transaction, the rules of inclusion and exclusion are the same, there are no conceptual differences (geographic, accounting, or due to Customs regime) and there are no errors (of coding or coverage). Included in "errors" are consistent interpretations of the classificatory schemes by one country which would be disputed by other countries or by the Customs Cooperation Council.

In principle, all sources of differences other than errors should be tractable although measuring the relative importance of different sources can be difficult in practice. A review of the different sources or factors is useful in order to consider how their effect can be accounted for in any comparison. Of these factors, transportation is probably the least difficult to deal with and almost certainly the least difficult to do something about. There are a number of countries such as the United States where imports are measured both ways: including and excluding transport. In principle, the information to estimate the cost of insurance and freight (c.i.f.) component across the board is available. Importers are legally bound to inform their Customs authorities of all their expenses in connection with a purchase abroad and the two broad categories of expenses are those that are dutiable (usually those connected with the product itself, including its packaging or wiring or mounting) and all others (usually those connected with the transportation, insurance and financing of the import). Accordingly, if it were necessary to conduct a study of transportation costs, there are administrative records which could be linked to the corresponding trade records. There are many technical problems related to how shipping and insurance information should be assigned to individual commodities in the case of complex shipments but there are proposals for ways to deal with these matters (Ryten 1983).

Equally, in principle, a study could be made of timing differences in the context of a particular flow of trade between any pair of countries. In the case of the reconciliation of trade statistics between the United States and Canada estimates were based on actual matches of documents which made it possible to compare dates and estimate average time lags between exports and corresponding imports. But there are less expensive methods to arrive at rough estimates that are also less constraining from the point of view of access to confidential records and are reasonably effective to calculate broad ranges of timing differences by points of exit and entry, by mode of transport, and by commodity.

Together, the estimates of timing differences and the difference between the cost of insurance and freight and free on board valuations (f.o.b.) can be expressed in the following equation:

$$A_m B(k) = B_x A(k) + A(\text{c.i.f.}) B(k) + \theta + e$$

where $A_m B(k)$ is the flow of imports for commodity k from country B to country A as recorded by Country A ; $B_x A(k)$ the counterpart flow as reported by country B ; $A(\text{c.i.f.}) B(k)$ the estimate of transport and insurance costs for that flow of trade as derived from country A 's records; θ a timing adjustment and e an error term that includes all the biases and random errors that affect both imports and exports statistics. It is assumed that all other sources of difference (geography, inclusions and exclusions, low value shipments etc.) have been disposed

of either by adjusting for them or preferably by excluding all transactions that may be affected by these factors from the comparison files. Over time, the average error should tend to zero and therefore the longer the period over which the comparison is made, the closer to each other the level or the average rate of change of the figures being compared. Should a comparison suddenly yield perverse results, this would constitute *prima facie* evidence of a deterioration in quality of at least one of the two terms of the comparison.

3.1 Analysis Using the World Trade Mini-Database

For purposes of analysis, a mini-database derived from the world trade database was created so as to start studying some of these effects. It covers the three principal trading blocs of the Western world: the EEC defined for these purposes as excluding Portugal and Spain; North America (Canada and U.S.A.); and Japan. Besides being simpler to use because of the reduced number of records, it avoids the problem of late reporting (mainly by third world countries) and of non-reporting (mainly by centrally planned economies). The mini-database includes exports and imports data for each of the constituent countries broken down by SITC (down to the four digit level of detail) and by partner country, from 1978 to 1985. In addition to the constituent countries, it includes two aggregates — the EEC and North America. Unlike the world trade database which includes a number of imputations to make analysis simpler, the mini-database only includes data as member countries reported them to UNSO after UNSO merged categories of trade deemed secret by the reporting country and converted non-standard codes reported by countries to standard SITC codes. None of these transformations is likely to affect the findings derived from the database in a significant way.

There are a few statistical problems with the grouping of countries in the mini-database. The United States has been reporting its imports to UNSO on the basis of c.i.f. but Canada reports imports f.o.b. Whereas the United States credits its partner countries on the basis of the origin of the imported goods, Canada reports on the basis of consignment (except for imports originating in Latin America). This in itself would not be too serious but for the fact that the United States is at times credited for exports routed to Canada. Accordingly, while the addition of the two countries should improve the matching of counterpart flows, the different systems of recording make it so much more difficult. Hopefully, this drawback will be overcome when United States f.o.b. imports are added to the base and when Canadian imports by origin replace imports by consignment for as many back years as possible.

In the case of the EEC countries, the key role that the Netherlands plays as port of entry to its European hinterland makes comparisons difficult. The Customs area of the port of Rotterdam acts not only as a giant distribution centre but also as a warehousing facility for the countries it serves. Accordingly, the exporter outside the EEC may not know to which specific country the sale is made but only that it will be warehoused in Rotterdam and for this reason credits the Netherlands with the sale. But the ultimate importer is bound by the rule of origin to assign the purchase to the correct country. As for the Netherlands, according to its records, no transaction involving goods has taken place across its Customs boundaries. It has simply sold harbour and warehousing services to either one of the transactors.

If the Netherlands served only the other members of the EEC as a port, the creation of an EEC total should suffice to improve the comparisons. But other countries (Switzerland and Austria in particular) also benefit from Dutch harbours and container terminals. This complicates matters somewhat because for example, the Swiss importer might apply the rule of origin to the Netherlands in cases where there has been a consolidation of imports from many origins. Or some value added operation performed outside the Customs zone in Rotterdam may not be reported as Dutch foreign trade in merchandise.

Another obstacle to interpretation is provided by the two Germanies — given that one fails to report its imports to UNSO and the other does not regard as exports the transactions it conducts with its Eastern counterpart. This means that there are extra-exports by the EEC that have no counterpart import records and, more specifically, that there are unreported trade transactions between the two Germanies. The size of this unrecorded leak varies with the relative affluence of East Germany and can only be surmised by looking at other indicators. There are also leaks that affect trade with Japan that will affect the results of comparisons involving Japan and its partner countries. These may be created by operations involving branches of Japanese firms located in S.E. Asia. However, the effect of these cases on aggregate data is not likely to be substantial and should not detract from the value of the analysis using this database.

i) Comparison of growth rates of counterpart statistics

Among the analyses conducted on the basis of the mini-database, one involved comparing growth rates in counterpart statistics, taking the period 1978-85. The assumption was that over that time period, the effect of errors and timing differences would be sufficiently attenuated so that the more permanent effects could be recognized. Moreover, by looking at growth rates, the effect of different valuations would be avoided to a considerable extent. The likelihood is small that the change in the cost of insurance and transportation is sufficiently different from the change in the average prices of the commodities transported to affect growth rates substantially over a period of three or four years. At least in the case of manufactured goods the proportion of transportation and insurance in the total cost is well below 10 per cent as borne out by United States ratios of f.o.b. to c.i.f. Moreover, transport costs would be only related to the weight and volume of the goods transported. Insurance costs, which are related to value, do not represent a significant proportion of total cost. And inter-transport mode substitution is unlikely to add to total cost in any other than exceptional circumstances. Accordingly, if the change in the corresponding cost were sufficient to affect import growth rates relative to counterpart export rates, the effects should be all in one direction and their size should vary with the average bulk of the commodities transported.

These speculations are only partly borne out by fact. Table 1 shows the differences in annual growth rates for counterpart total trade for the pairs of origins and destinations derived from trade among the EEC, North America, and Japan. While relatively small, these differences do not suggest any pattern though there may be some underlying regularities that escape superficial inspection.

Table 1
Differences in Growth Rates for Counterpart Annual Total Trade
for Japan, North America and the EEC, 1978-1985

Country A - Country B	Difference in growth rate for the period ¹			Difference in value of exports in 1982 in millions of dollars ²
	1982/78	1985/82	1985/78	
N.A. - EEC	.6	-.5	-.5	265
N.A. - Japan	-.4	.5	-	-
EEC - N.A.	-.8	-.7	-.8	365
EEC - Japan	1.1	1.9	1.5	90
Japan - N.A.	-.7	-.2	-.5	200
Japan - EEC	-1.2	-.6	-.9	155
Mean Absolute Difference	.8	.7	.7	

¹ Defined as percent growth in $A_x B$ less percent growth in $B_m A$.

² Difference between $A_x B$ and $B_m A$ rounded to the nearest five million dollars.
A dash (-) denotes an insignificant value.

Table 2
Differences in Growth Rates for Counterpart Annual Total Trade by SITC Section
Japan in 1978-82 and 1982-85

SITC Section	Japan - North America			Japan - EEC		
	Difference in growth rate for the period ¹		Difference in value of exports in 1982 ²	Difference in growth rate for the period		Difference in value of exports in 1982
	1982/78	1985/82		1982/78	1985/82	
5. Chemicals	.7	-1.6	15	-1.5	.5	5
6. Semi-manufactures	-2.5	.9	60	1.9	-.5	5
7. Transportation equipment	-1.0	-1.0	275	-2.0	-.7	85
8. Miscellaneous manufactures	1.4	-.9	35	-.8	.8	20

¹ Defined as percent growth in $A_x B$ less percent growth in $B_m A$ (A is Japan).

² Difference between $A_x B$ and $B_m A$ in millions of dollars rounded to the nearest five million dollars.

Table 2 shows growth rates for selected SITC Sections between Japan and its two trading partners. The principle involved in simplifying Table 2 was to ignore flows with less than one million dollars in 1982 since such flows do not appear to be sufficiently stable to warrant interpretation.

Discussions about internationally comparable commodity classifications have invariably demanded more rather than less detail. The collection of statistics for purposes of international comparison has induced countries to publish data well beyond the 3-digit of the SITC or its equivalent. A number of third world countries publish data broken down by ten digits corresponding to nationally-annotated international classification, and country. Inspection suggests that flows coded at one digit — where there has seldom been any controversy — are subject to very considerable differences when compared with their counterparts as soon as their absolute value drops to, say, below 50 million dollars. Beyond the first digit of the classification, differences rise very rapidly.

The case of Japanese exports to North America and counterpart imports shown in both tables 1 and 2 warrants further consideration. At mid-point (1982) this trade was valued at about forty billion dollars (US). Total imports grew on average by half of one per cent per annum more than exports. This is an amount of about two hundred million dollars per annum at mid point. Detailed examination suggests that a substantial part of the explanation lies with section 7 of the SITC which includes inter alia all types of transport equipment. There the difference in growth rates is of one per cent per annum on average. It would be interesting to pursue this investigation to determine whether the discrepancy is evenly distributed or whether its incidence is chiefly felt by one particular commodity.

But whatever the causes, these comparisons suggest that over a sufficiently long number of years and for comparatively large portions of total trade flows, differences in growth rates are not large in absolute terms. Notwithstanding this observation, even small differences could play havoc with period-to-period changes in the overall trade balance, particularly when it is close to zero. Moreover, when dealing with a trading partner such as Japan, with exports heavily concentrated in one or two one-digit breakdowns of the commodity classification, the possibilities of compensation for systematic misclassification are comparatively few. This makes it all the more important to understand why bilateral trade as measured by the two counterpart reports has not been moving in step.

Table 3
Changes in *X/M* Ratios Between 1978 and 1985 and Comparisons with Standardized *X/M* Ratios Assuming Constancy of SITC Section Shares¹

	North America			EEC			Japan		
	Simple Ratio		Std Ratio	Simple Ratio		Std Ratio	Simple Ratio		Std Ratio
	1978	1985	1985	1978	1985	1985	1978	1985	1985
North America				.90	.91	.92	.85	.86	.69
EEC	.96	.92	.69				.78	.86	.89
Japan	.95	.98	.91	1.00	.94	.86			

¹ The simple ratio is $X/M = (A_x B/B_m A)$. The standardized ratio using common shares is

$$\text{Std ratio} = \frac{1}{M_{78}} \sum_{i=0}^n \frac{x_{it}}{m_{it}} \cdot m_{i78},$$

where m_{it} = current imports for section i of the SITC ($i = 0, 1, \dots, n$),
 m_{i78} = imports in 1978 for section i of the SITC,
 x_{it} = current exports for section i of the SITC, and
 M_{78} = total imports in 1978.

ii) *Comparison of the ratios of annual exports to imports*

A different kind of analysis was also very revealing. Any import flow should be equal to the counterpart export plus the cost of freight and insurance plus some term which reflects the sum of conceptual differences, timing, and errors. Whereas timing and errors should make their impact felt mostly in the short term, conceptual differences should emerge as the dominant influence in the longer term. For this reason, if the ratio of annual exports to annual imports changes over time this can be due to a combination of the following factors: because of a change in the shares of relatively high c.i.f. to low c.i.f. components; because of a change in the mix of commodities with small relatively to commodities with large-timing differences; because of a change in the proportion of c.i.f. to total value; and because of other factors.

Table 3 shows some aggregate results of this analysis. Against each of the flows involving Japan, the EEC and North America, there are three figures: the simple (current year weighted) ratio of aggregate exports to aggregate imports in 1978, the corresponding ratio in 1985 and the standardized base year weighted ratio assuming that the proportions of imports by section to total imports for each flow of trade remained constant since 1978. These standardized ratios are an approximation to an estimate that removes the impact of variations in the mix of c.i.f. from the variation in the ratio over time. Any difference between the 1978 and the standardized 1985 ratios should therefore be ascribed to other factors.

There are expectations about the way ratios should change over time as a result of the increased share of highly manufactured goods in certain export flows. For example, exports by the EEC to North America and Japan; exports by Japan to the EEC and to North America can be expected to include proportionately more manufactures. Accordingly, the ratio that reflects changes in mix is higher than the standardized ratio. This follows because the relative importance of c.i.f. decreases as the value of a unit of weight or volume increases.

But there are *a priori* exceptions to this prediction shown up by the table. For example, the exports of North America to Japan show a very large gap between the simple and the standardized ratios even though the share of manufactures went up relatively less.

Table 4

Variations in Simple x/m Ratios Between 1978 and 1985 Compared with Standardized x/m Ratios with Constant SITC Division Shares

Exports from ... to ...	N.A. EEC	EEC N.A.	EEC Japan	Japan EEC	N.A. Japan	Japan N.A.
SITC Sections						
0 Food	107	102	100	98	98	90
1 Beverages & tobacco	99	100	99	100	99	100
2 Crude materials	100	100	96	93	100	102
3 Mineral fuels	102	117	304	93	103	109
4 Animal & veg. oils	99	98	107	100	101	92
5 Chemicals	102	101	101	97	100	96
6 Manufactured goods	99	101	99	96	98	100
7 Machinery & transport	96	91	95	97	92	100
8 Misc. manufactures	100	100	100	98	97	99
9 Misc. transactions	150	176	163	157	86	92

Table 4 provides a breakdown by SITC sections for the ratios corresponding to trade flows between each of six pairs of trading blocks recorded in the mini-data base. The figures shown are ratios of the simple index at the Section (1-digit) level to the index derived using share of imports at the Division (2-digit) level. They indicate the contribution to the variation in ratios accounted for by changes in the commodity mix. They are no more than indicators partly because they only go down by one level in the commodity classification.

(Figures in the table are derived by taking the index that measures the change in each section of the simple X/M ratio from 1978 to 1985, i.e., (x/m) 1985 divided by (x/m) 1978 and dividing it by a corresponding index in which the standardized (x/m) ratio for 1985 was used and where the division ratios were aggregated using their 1978 shares in their corresponding division. Simple algebra suggests that the ratio obtained R_i is:

$$R_i = 100 \cdot M_{i78} \cdot \frac{X_{i85}}{M_{i85}} \div \sum_{j=0}^{n_i} \frac{x_{ij85}}{m_{ij85}} \cdot m_{ij78}.$$

Notation is similar to that used in table 3. Subscript i denotes the section and subscript j denotes the division within the section ($j = 0, 1, \dots, n_i$). A figure of 104 for example implies that a four percent increase in the current value of exports relative to counterpart imports took place for reasons other than the effect of changes in commodity mix on the c.i.f. component.)

No pattern is readily detectable: there are roughly as many cases which overshoot as cases which undershoot the mark. For the bigger flows, such as North America to EEC or EEC to Japan, the commodity mix is relatively stable as a result of which there is little difference between base and current weighted ratios (except for those sections of the SITC where trade is comparatively small as in the case of Mineral fuels exported by the EEC to Japan). Moreover, these do not move that much over the period. Other flows are very sensitive to the commodity mix which suggests that at lower levels of the classification f.o.b./c.i.f. differences explain a small portion of the variation in x/m ratios over time.

3.2 Analyses Using the Complete World Trade Data Base

Potential country and commodity mis-classifications:

Tables 5 and 6 derived from the complete world trade data base present counts of potential country and commodity misclassification. Table 5 presents a count of the number of cases in 1983 in which there is bilateral trade in a commodity according to one of the reporting countries of a trading pair but not according to the other. This is shown for each level of SITC detail as a proportion of all cases. Table 5A shows the impact on value, again for each level of the SITC. In addition to providing a summary measure of the size of errors, the tables also give an idea of how fast the number of anomalous situations increases as a function of the detail of the classification.

Table 5
Comparison of Foreign Trade Statistics in 1983 – Number of Records¹

SITC Level of Detail	Percentage Reporting No Exports	Percentage Reporting No Imports	Total Percentage
0 (overall)	11	4	15
1 digit	14	7	21
2 digit	16	10	26
3 digit	19	13	32

¹ Percent of number of records of trading pairs with one member reporting no exports/imports while other member reports non-zero trade.

Table 5A
Comparison of Foreign Trade Statistics in 1983 – Value of Records¹

SITC Level of Detail	Percentage Reporting No Exports	Percentage Reporting No Imports	Total Percentage
0 (overall)	.1	–	.1
1 digit	.3	.1	.4
2 digit	.6	.4	1.0
3 digit	1.1	.9	2.0

¹ Percent of value of records of trading pairs with one member reporting no exports/imports while other member reports non-zero trade.

A dash (–) denotes an insignificant value.

Table 6
Comparison of Counterpart Foreign Trade Statistics in Two Selected Years

	1979	1983
Number of records with $x > m$ as percent of all records	35	32
Value of exports where $x > m$ as percent of total exports	41	42
x/m ratio for $x > m$	1.18	1.15
x/m ratio for $x < m$.87	.85

Table 7
X/M Ratios in 1985
 From Three Selected Reporting Countries to Nine Trading Partners

To	From	Canada	U.S.A.	Japan
E.E.C.		.84	.92	.94
Netherlands		1.93	1.34	1.33
Belgium - Luxembourg		1.47	1.51*	1.26
Denmark		1.20*	.74	1.05*
France		.70	.74*	.69*
Germany, F.R.		.69	.81	.98
Ireland		.55	.78	.72*
Italy		.75	.84	.75*
U.K.		.74	.86	.89
Greece		1.00	1.23*	.89*

Table 6 shows changes between two selected years in a number of indicators — related to cases where exports are in excess of counterpart imports. While over a period of four years there has been some change in the percentage of records for which exports exceed imports as well as in the percentage value of total exports for those records, the changes in question are minor. Surprisingly, the cases of x/m account for more than 40 per cent of the total value of trade and as this figure went up fractionally, the proportion of records that accounted for it fell by 10 per cent.

In the case of Table 7 a number of *a priori* predictions are tested against fact. Three reporting exporters — Canada, United States and Japan — and nine reporting trading partners — the members of the EEC other than Spain and Portugal are studied. The tables list the 1985 simple x/m ratios for country to country trade. Other things being equal, the following predictions seem plausible:

- the higher the manufacturing content of a trade flow, the higher the x/m ratio, which is equivalent to saying that the c.i.f./total value ratio is smaller, the more value added is embodied in a commodity. For this reason, the ranking in *ascending* order of ratios should be Canada, United States, Japan;
- in the case of trade with the *entrepôt* countries — Netherlands, and to a lesser extent, Belgium Luxembourg — country miscoding by the exporter should apply mostly to bulk shipments. For this reason the x/m ratio in *descending* order should be Canada, United States, Japan; and
- x/m ratios greater than one should only occur for *entrepôt* countries.

For thirty x/m ratios (counting in the three ratios for the EEC as a whole) there are nine cases (entries with * in table) for which the predictions do not hold. Removing Greece's two because the corresponding trade flows are much too small, seven ratios do not behave according to expectations which is still in excess of twenty percent of all cases.

The critical finding in these analyses is that any increase in the level of detail in the classification hierarchy beyond the combined one makes comparisons with counterpart trade very difficult. This is not compatible with the progressive attempts, conducted both nationally and internationally, to expand the detail of the commodity classification and to increase the number of breakdowns by additional classification variables. Even when pooled over time, the transactions in these detailed cells match poorly with their counterparts. Since it cannot be argued that both reports involved in a bilateral comparison are simultaneously correct, the chances are that both contain a significant error component.

4. MAKING USERS AWARE OF ERROR

There are two separate issues. One is to make users aware, that contrary to widespread belief, the foreign trade figures, particularly the detailed figures, may be flawed. The other is to put together a programme to improve the quality of foreign trade data taking advantage of the fact that counterpart measurements of the same transaction exist. A number of proposals to get such a programme underway follow.

The analysis presented in this paper provides that beyond the two-digit level of the commodity classification by country, even annually, neither levels nor year-to-year changes can be taken with complete confidence. Users will probably not take kindly to such a finding, as they already have reason to question the coverage of aggregates in the case of exports. The results of the reconciliation programme between the United States and Canada should not be viewed as limited to the two countries. Others experience the same class of problems to a varying extent. The revelation that, in addition to these weaknesses, data by commodity beyond a certain level can only be used with great caution, could lead to a fundamental change in the perception that users have of foreign trade statistics.

But, if this measure is not taken, no matter how unpopular the news, a belief that has less than full underlying factual support is perpetuated. The detailed commodity figures are used in a variety of ways and the one that is most topical is for purposes of tariff policy. Discussions on these matters rely heavily on detailed figures, seldom on the differences between national and counterpart data, and equally seldom on domestic consumption statistics as a check on the orders of magnitude suggested by Customs data. Moreover, in another use of detailed commodity data, views about industrial and regional policy are formed and actions may be taken on the basis of evidence which this analysis suggests is not solid. Surely it is incumbent on statistical agencies to make users aware of the perceived inadequacies of the data in order to prevent the generalization of their misuse.

5. A PROGRAMME TO IMPROVE FOREIGN TRADE STATISTICS

In addition to providing users with more factual information about error in foreign trade statistics, a programme or programmes to improve the quality of these statistics over time should be formulated. The following are steps which should probably have been taken some time ago:

- i) the c.i.f. component of imports should be measured systematically. Without it, it will not be possible to compare exports with imports across the board. The information is available at the time the import is reported to Customs. Matters such as how often and to which detail will depend on resources and on the urgency to improve the knowledge of users;
- ii) an inquiry should be launched into time lags between exports and imports by commodity category and by country of origin. To make such a study effective, it is probably necessary to count on the co-operation of partner countries; although, if this is not forthcoming, reference to commercial invoices may be an acceptable surrogate;
- iii) on the basis of knowledge of these two elements, a formal method to estimate counterpart imports on the basis of exports should be used and the error of estimate tabulated for future study. If the error of estimate has no significant autocorrelation properties, coding and related errors might explain the difference between the recorded import and its statistical estimate. If, however, the error term does not satisfy these criteria, it should be marked down for future inquiry in co-operation with the partner country;

- iv) obvious surpluses or deficits should be tested against countries likely to play the role of commercial intermediary or entrepôt. For example, an export surplus with the Netherlands for the United States should be tested against corresponding deficits with such countries as the Federal Republic of Germany or France. Econometric methods can be used to disentangle an across-the-board effect of entrepôt services (although they are more likely to be used for bulky and warehousable merchandise) from short-lived effects such as coding error;
- v) for those commodities which are systematic outliers, after all adjustments have been made, either because they persist over time or because they occur across countries, advantage should be taken of the Harmonized System by enlisting the help of the Customs Cooperation Council for the interpretation of its explanatory notes.

Obviously the launching of such a programme requires preparation, approval, and resources. It cannot take place at once nor will it be sponsored by most countries straight away. But the proposals ought not to be shelved as similar proposals were some thirteen or fourteen years ago. There is too much attention paid to the trade statistics to risk delaying their improvement. Their comparison with counterpart data shows that they can only stand increased attention if they are substantially improved or if their analysts become more aware of the limitations of the material on which they test their hypotheses.

REFERENCES

- ALLEN, R.G.D., and ELY, J. EDWARD (1953). *International Trade Statistics*, New York: John Wiley & Sons.
- COATS, R.H. (1926). Canadian Trade Statistics. *Journal of the Canadian Bankers' Association*.
- RYTEN, J. (1983, 1984, 1986). Reports on the Reorganization of Bolivian Foreign Trade Statistics (In Spanish). United Nations Development Programme. New York.
- UNITED NATIONS, (1982). International Trade Statistics, Concepts and Definitions. *Statistical Papers Series M*, 52. New York.
- UNITED NATIONS (1986). Standard International Trade Classification. *Statistical Papers Series M*, 34. New York.
- UNITED NATIONS STATISTICAL COMMISSION (1974). International Trade Reconciliation Study. Report of the Secretary General. Eighteenth Session, Geneva.

Multipurpose Sample Designs¹

LESLIE KISH²

ABSTRACT

Most surveys have many purposes and a hierarchy of six levels is proposed here. Yet most theory and textbooks are based on unipurpose theory, in order to avoid the complexity and conflicts of multipurpose designs. Ten areas of conflict between purposes are shown, then problems and solutions are advanced for each. Compromises and joint solutions fortunately are feasible, because most optima are very flat; also because most "requirements" for precision are actually very flexible. To state and to face the many purposes are preferable to the common practice of hiding behind some artificially picked single purpose; and they have also become more feasible with modern computers.

KEY WORDS: Allocations to domains; Mean-Square-Errors; Multipurpose allocation; Multipurpose design; Optimal allocation; Periodic samples; Sample size.

1. INTRODUCTION

Most studies involve several purposes during the planning stages and then typically many more purposes emerge later during the analyses of data and more during their interpretation and utilization. However, the real multipurpose nature of most studies tend to remain hidden under the surface of oversimplified, univariate discussions of study designs. This seems most clearly evident for sample surveys, which I shall discuss here; but I believe that this discrepancy also holds for other statistical designs, such as experimental and evaluation studies.

In practice, surveys are usually multipurpose. Why then are multipurpose designs neglected in sampling theory? Because multipurpose theory would be too complex and difficult, and sampling theory is rather complex already; specific exceptions will be noted later. Even the descriptions we read of actual sample designs tend to follow and to borrow the prestige of univariate and unipurpose sampling theory, rather than to portray faithfully the many compromises of complex reality. Many common designs (especially equal probability of selection method) probably serve robustly a variety of purposes, *explicit* planning of multipurpose designs seems to be rare, though much needed, I propose.

There are several aspects to the *multipurpose nature* of survey samples, and these are displayed in a hierarchy of *six levels* in Section 2. Then *ten areas of conflict* between purposes are specified in Section 3. Sections 4 to 9 deal with specific areas of conflict, presenting approaches to and solutions for them. Some of these solutions are attributed to widely dispersed articles of survey sampling; but others are more novel, hence less fully developed, derived, and referenced.

In this overall review I aim first and foremost to serve practitioners with handy references on approaches, methods and procedures for multipurpose designs; to alert them both to the importance and to the feasibility of such designs. Second, I also wish to provide a framework for integrated, theoretical future work on the many problems and conflicts of multipurpose designs. Imperfections of my methods can serve as stimuli to others for better derivations for them, as well as for developing new methods.

¹ Keynote address at the International Symposium on Statistics, Taipei, Taiwan, August 1986; and also at a seminar of Statistics Canada, October 7, 1987.

² Leslie Kish, Institute for Social Research. The University of Michigan, Ann Arbor, MI 48104 USA.

2. A HIERARCHY FOR LEVELS OF PURPOSES

To begin with, we need some clarification of the meaning of "multipurpose", because too many concepts are confused under this term in our literature. To reduce the confusion, I classified a score of purposes into six levels in Table 1. Most of the time either multiple variables or multi-subject surveys (levels 3 or 4 in Table 1) are discussed and "multi-subject" (4) has sometimes been distinguished from multipurpose (3) for the same or closely related variables (Murthy 1967). Each of these six levels is shown in several specific manifestations, which can be usefully augmented and discussed in more detail elsewhere (e.g., United Nations 1980; Lahiri 1963).

Integrated survey operations on level 5 are related to, but should be distinguished from multi-subject surveys, because they refer to organizations and institutions that conduct many surveys in diverse fields over longer periods of time (United Nations 1980; Foreman 1983). An earlier name was "continuing survey operations", when it was recognized that most large-scale, wide-spread sample surveys were conducted by continuing survey organizations like the U.S. Census Bureau, Statistics Canada, or our Survey Research Center. Such continuity has large advantages in costs and quality, with restraining effects on sample designs (Kish 1965).

Master frames or master samples on level 6 refer to further extensions and specializations of multipurpose approaches. They may refer simply to using the same maps, or block listings, or area segments for several different surveys; or to the large-scale example of the "Master Sample of Agriculture" (King and Jessen 1945), where rural areas on the maps of all the counties of the USA were divided into segments of about four farms each; or to the firm that sells current listings of dwellings for most samples used in Western Germany. These very diverse examples have common bases in the savings from sharing the "startup" costs (of design, stratification, listing, etc.) for constructing sampling frames.

Diverse statistics based on single variable and diverse domains (levels 1 and 2) have been typically neglected in the literature of multipurpose sampling, although they are the most common, but they can have the most drastic effects and cause the most dramatic conflicts, as we shall see later. The effect of designs can be very different for statistics like medians and quantiles or regression coefficients than the effects for means and for aggregates (Kish 1961; Kish 1965; Kish and Frankel 1974). Furthermore, designing for period samples brings on new considerations (Section 8). But most dramatic effects can be seen simply for the means of small "subclasses" (e.g., as small as 0.10 or 0.01) of the entire sample, representing similar "domains" in the population (Section 5).

Each of the six levels of purposes presents different aspects for designs and each level can be fruitfully explored for more specific meanings and examples, some of which are listed in Table 1.

The difficulties of multipurpose designs, which have caused them to be neglected and avoided, are of several kinds. First, the different purposes must be formulated *explicitly in statistical terms*, so that these may serve in formulas for their comparisons and for formulated compromises; but obtaining a (complete) list of such explicit, formal terms may be the principal obstacle. Second, estimates of *variance and cost factors* are needed for each purpose. Third, for some methods values must be obtained for the assigned to the "*required*" *precisions* for all the purposes (Section 5). Fourth, the above values and estimates must be combined in a mathematical formulation in order to arrive at the *solution of a single "optimal" design* to be actually used. The computational tasks for such solutions have been eased by electronic computers, but the conceptual and theoretical tasks remain (Section 5).

The difficulties of these tasks help to explain why discussions of multipurpose designs have been largely neglected designs in textbooks. However, note later references and bibliography here and in Rodriguez-Vera (1982); also Cochran (1977), and Chatterjee (1967). Furthermore, also in descriptions of actual surveys, often a single statistic (e.g. the mean) of a single principal variable is presented as *the* only (principal) purpose for the study. In the framework of multipurpose design

Table 1
Hierarchy of Purposes

-
1. Diverse statistics from the same variables
 - Totals or means or medians and quantiles, distributions
 - Analytical statistics: regressions, categorical analysis
 - Time aspects: static, macro-change, micro-change, cumulative
 2. Diverse populations and domains (subclasses)
 - Proper classes and crossclasses
 - Comparisons of subclasses
 3. Multiple variables on the same subject
 - Alternative measures of one variable;
e.g. of income, or unemployment
 - Diverse periods — per day, week, month, year
 - Several aspects of one subject: income, savings, wealth
 4. Multisubject surveys
 - Several subjects on same schedule, interview, operation
 - Health surveys of many diseases
 - Market research for several clients, many goods
 - Agricultural surveys of many crops
 - "Omnibus" social surveys
 5. Continuing, integrated survey operations
 - NSS in India, CPS in USA, NHSCP of UN
 - Separate surveys from one office and field staff
 - Common source of surveys
 - Diverse methods, costs, operations, allocations, respondents
 6. Master frames
 - Several samples from one frame or set of listings
 - Separate institutions, organizations
 - Separate field staffs? Same PSU's?
-

design this is equivalent to assigning zero importance to all other purposes. The unreality of this pretense may be softened by assuming that other principal purposes would result in similar allocations; but this pretense should be buttressed with calculations of the four steps above.

3. AN OVERALL VIEW OF TEN AREAS OF CONFLICT

A brief overall view of ten areas of conflict, listed in Table 2, should be useful before we look at specific problems and possible solutions for each. The list will probably not prove exhaustive, and readers may well find other areas. Even more likely, they may find within these ten areas other problems and other solutions not explored here. It would be convenient if the ten areas of conflict should be linked rationally to the twenty purposes presented in six levels; we then could reduce this presentation to say, twenty purpose/conflict nodes or to ten level/conflict nodes. Unfortunately the areas of conflict denote a perpendicular dimension to the purpose and all (or most) of the 10×6 cells have meaningful contents.

Of this long list of ten areas of conflict fortunately not all need to be formulated for every actual sample design. I believe that possible conflicts about a) the sample sizes m_g and about b) the relation of biases to sampling errors should always be considered, at least informally, because they

are ubiquitous. Also c) allocation among domains and d) allocation among strata should receive at least a brief discussion, and often more. Computing sampling errors (j) should also be done on most surveys. However, in the common case of one-time surveys, conflicts i) about design over time need not be considered. On the other hand, in a continuing operation with a continuing sampling frame, the decisions about e), f), g), and h) (stratification, cluster sizes and measures) may have been made a long time ago for a fixed design. However, the cluster sizes (f) used in intermediate stages (blocks and segments) may be open to flexible operational changes.

It is also reassuring to know that compromises based on statistical methods can yield quite acceptable results, for several reasons (Sections 5-8). First, because moderate departures from optimal allocation result in only small or negligible increases of variance. Curves of efficiency tend to be flat within broad areas around the optimal points; thus great accuracy for separate designs, which would not be feasible, are not needed. Second, because wide departures from optimal allocations can, on the other hand, cause moderate to large increases in variances. Thus, ignoring important purposes can result in substantial losses of efficiency for them, and therefore those purposes should be included in compromise designs. Third, compromise designs, in accord with statistical methods, can reduce drastically the potentially large losses from allocations optimized for other purposes, and with only small increases over the separate optimal designs for each purpose (Section 5).

4. SAMPLE SIZES AND BIAS RATIOS (B/σ)

These two areas of conflict, a and b in Table 2, should perhaps be considered most important overall, because they can be most dramatic. We treat them together here only because they may be closely related through the effects of subclasses. Let us begin with the familiar (simple random sampling with replacement) sample size $m = S^2/V^2$ needed to yield a "required" precision $= V^2$ for a sample mean \bar{y} , with element variance $= S^2$. However, the S_g^2 depend greatly on the variables and on the domains, indexed jointly with g for the year \bar{y}_g ; and the "required" V_g^2 may vary even more. We also include design effects D_g^2 that also vary, and thus $m_g = S_g^2 D_g^2 / V_g^2$ expresses the sample size needed for the mean of the variable g . For the mean \bar{y}_g of a domain g , comprising only the proportion P_g in the population the *overall* sample size needed for the domain becomes $n_g = m_g / P_g$, and it is more practical to formulate the needed sampling fraction $f_g = n_g / N = S_g^2 D_g^2 / V_g^2 P_g N$. The factor $(1-f)$ may be neglected or included in D_g^2 . The P_g become small and critical if high precisions are "required" for small subclasses.

For comparisons of subclasses the variances increase even more: $m_g = (m_a^{-1} + m_b^{-1})^{-1} = n(P_a^{-1} + P_b^{-1})^{-1}$, with the P_a and P_b denoting proportions in the sample n (assuming $S_a^2 = S_b^2$). E.g., for the comparison of two subclass means of $0.01n$ and $0.10n$, we have the "effective size" $m_g = n(0.01^{-1} + 0.10^{-1})^{-1} = n/110$. For other statistics, such as medians and regression coefficients, formulating "required" sample sizes would become complex. It is more than we may discuss here, but some numbers may probably be specified.

Considerations for subclass statistics become greatly modified if, in addition to variances σ^2 , we also include biases B^2 in the Root-Mean-Square-Error $= \text{RMSE} = \sqrt{(\sigma^2 + B^2)}$ for measures of accuracy. Figure 1 is meant to portray a common tendency in the accuracy of survey data, although great differences in the relations of biases to sampling errors are possible; reading the legend is urged here. It occurs commonly that potential biases B_1 are greater than the measurable sampling and variable errors σ_1 , for the entire sample. However, on the horizontal axis the standard error σ_1 is shown to increase by a factor of about 3 for σ_2 of a subclass of about 1/10 of the total sample. For comparisons (differences) of two such subclasses σ_3 increases by about 1.4 more.

Table 2
Ten Areas of Conflicts (a-j)

a.(4) ¹	Sizes m_g or rates f_g are needed for purposes g $V_g^2 = S_g^2 D_g^2 / m_g \text{ and } m_g = S_g^2 D_g^2 / V_g^2 \text{ or } f_g = S_g^2 D_g^2 / V_g^2 P_g N$ where m_g denote subclass sizes and $f_g = n_g / N = m_g / P_g N$ denote sampling rates
b.(4)	Relation of biases to sampling errors in $RMSE = \sqrt{(\sigma^2 + B^2)}$ - The bias ratio B/σ decreases as σ increases for subclasses - For comparisons B/σ tends to be small as B decreases, σ increases
c.(5)	Allocation of the m_g among domains $m_t = \Sigma_g m_g$
d.(6)	Allocation of m_{gh} among strata h $m_g = \Sigma_h m_{gh}$
e.(6)	Choice of variables for stratification Multivariate stratification
f.(7)	Optimal cluster sizes $D_g^2 = [1 + \rho_g (\bar{b}_g - 1)] \bar{b}_g = P_g n_t / a \text{ for crossclasses}$
g.(7)	Measures for cluster sizes
h.(7)	Retaining sampling units (PSU's) for changed subjects, measures and strata and for diverse subjects.
i.(8)	Design over time How much overlap? Panels? Change versus cumulation.
j.(9)	Computing and presenting sampling errors.

¹ The numbers (4) to (9) refer to sections with treatments.

However, the hypotenuses denoting the RMSE are shown to increase much less. In $RMSE_1$ the bias B_1 is shown to dominate, and this may happen for some variables in large total samples. However, the subclass $RMSE_2$, because the bias was kept constant at $B_2 = B_1$, increased only moderately and is dominated by σ_2 . This is even more true for $RMSE_3$, where the σ_3 has increased, but the biases — assumed to have the same sign, because that is a common tendency — decrease B_3 in the difference of means.

Examples of these phenomena abound everywhere and for all purposes are listed in Table 1. We choose the best known, critical statistics of unemployment, where admitted measurement biases may completely swamp the low values (e.g., 0.1 percent) of measurable fluctuations. However, for small subclasses (e.g. Black teenage boys) the sampling errors for small sample bases overtake the biases. For periodic comparisons the sampling variations become even more critical.

These relations among biases and variable errors assumed here are not logically necessary, but empirical and common. Neglect of these simple relations leads to a great deal of confusion concerning the need for sample surveys of adequate precision, i.e. with small sampling errors, σ . I propose Figure 1 as practical answers to some common questions, such as: Why do we spend

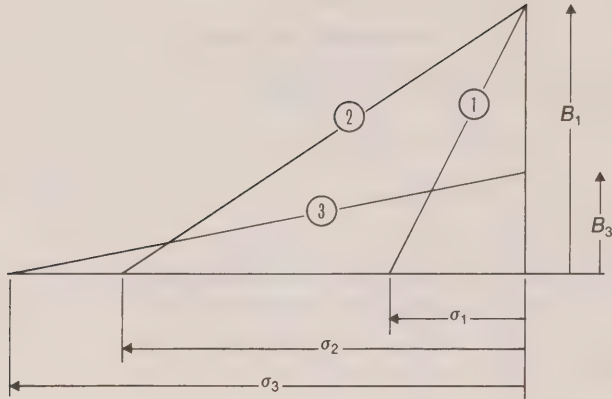


Figure 1. Variable errors (σ) and biases (B) in root mean square errors (RMSE)

The bases represent sampling errors and other variable errors (σ). For example σ_1 may be the $ste(\bar{y}_l)$ for the mean \bar{y} of the entire sample and σ_2 may be a larger $ste(\bar{y}_c)$ for a subclass mean, and σ_3 may be the $ste(\bar{y}_c \text{ } \bar{y}_b)$ for the difference between two subclass means.

The heights represent biases (B) and the hypotenuse denotes the $RMSE = \sqrt{(\sigma^2 + B^2)}$. (1) For the entire sample the bias B_1 may be large compared with the variable error σ_1 , thus taking larger samples would not decrease the $RMSE_1$ by much. (2) However with the same bias B_1 , but with a smaller sample in the subclass, the ratio changes and the σ_2 dominates the $RMSE_2$; and this is not much larger than for (1) despite a much smaller sample. (3) Furthermore, for the difference of means, the net bias B_3 may be much smaller; so that even with a larger σ_3 , the $RMSE_3$ for the difference is but little greater than $RMSE_2$. This drastic change in the bias ratio B/σ tends to appear not only for differences between subclasses within the same sample, but also for differences between repeat surveys.

money for large samples and on rigorous sampling methods in the face of large measurement biases? Why bother computing sampling errors when response biases dominate the total error? The implicit answers come from the domination of sampling errors in the subclasses, and even more in their comparisons. Let us make these implicit answers more explicit in future sample designs.

5. ALLOCATION AMONG DOMAINS

This most important and frequent area of conflict has several aspects. First, consider the allocation of total sample size (or effort or cost) among the domains that constitute a partition of the total population. A common example is allocation among the several (5, 10, 20 or 50) provinces or regions or states of a country; those domains typically have very unequal populations N_d , with ranges of 1 to 100 perhaps in relative sizes, though they may cover roughly equal surface areas. Often the question takes this form: Should the sample sizes n_d be roughly equal; or should the n_d be proportional to the N_d , with constant sampling rates $f_d = f$? Equal n_d tends to yield roughly equal errors, $ste(\bar{y}_d)$ for the means. On the other hand, constant $f_d = f$ tends to yield the lowest $ste(\bar{y}_w)$ for the overall mean $\bar{y}_w = \Sigma W_d \bar{y}_d$, because it yields lower errors for the larger domains. This error may be lower than “required” for \bar{y}_w , especially in view of potential biases (Figure 1), and may not justify large total sample sizes and costs. This is the contention of proponents of equal sizes n_d for provinces. However, increased sampling errors for \bar{y}_w are also suffered by most other subclasses, especially “crossclasses” like age, sex, socioeconomic classes, etc. whose sizes tend to proportionality to the total. Those are common disadvantages of the highly unequal $f_d = n_d/N_d$ for provinces that result from the equal n_d values.

For example, in the Current Population Surveys of the USA, larger f_d are assigned to the smaller states. The resulting weighting increases the variances (for a fixed total cost) of the overall means and also of “crossclasses”, such as young men and women, and especially of Black teenage

boys and girls (with critically high unemployment rates). Similar conflicts between national and provincial needs occur in all countries, because provinces have widely different populations. The need for better provincial data, for fixed total cost, conflicts with greater precision for national and for "crossclass" statistics.

To reduce the usual confusion, I distinguish "domains" to denote partitions of the population, from "subclasses," the corresponding partitions of the sample. Then I distinguish "design domains" (and subclasses) to refer to partitions (like provinces and regions) that are contained in strata defined by the sample design, from "crossclasses" (like age, sex, occupation, income, etc.) that cut across the sample design, both clusters and strata, often almost randomly. The design effects differ for these two types of subclasses (Kish 1961, 1980, 1987).

In addition, other sources of conflict may arise from *domain* differences other than their sizes: in the distribution of variables, also in the variances $D_d^2 S_d^2$ precisions; but we need not enter into those complexities here. Beyond calling attention to the problems, we refer to two distinct technical methods for the joint solution of the conflicts in allocation, (the fourth step noted at the end of Section 2). One approach uses iterative nonlinear programming in order to satisfy for *minimal cost* the "required" precisions jointly for all stated purposes. These elegant solutions to diverse problems exploit modern computers and have been published in many articles since 1963 (see reviews and references in Bean and Burmeister 1978, Rodriquez-Vera 1982, Cochran 1977). The "required minimal" cost often turns out much too high, because the "required" precisions were unrealistic. Then the solutions are drastically rescaled downwards. But such rescaling exposes the false pretensions (in my view) of this elegant approach that depends on unrealistic "required" precisions. Principally, I question the reality of "step functions" for "required" precisions that assign a constant value to any variance below the required V^2 and zero value to variances above it.

A very different approach calls for some form of *averaging* between all the "optimal" (preferred) allocations for various purposes, by *minimizing the combined* (weighted) *variance* either for fixed cost or fixed sample size. Of course, if the resulting combined variances turns out to be too high (or low), the solutions can be scaled up (or down) in total fixed cost or sample size. I prefer this solution, which compromises between different allocations, each of which would optimize for only one purpose (Yates 1981; Dalenius 1957). It involves assigning relative values of importance I_g to all the list statistics and this may seem difficult (but an "ignorant" decision-maker can assign equal I_g to all of them). But the other two alternatives are more extreme and they are bound to prove even more difficult: either to specify the "required" precisions of all statistics for the first approach, which then assigns arbitrarily equal weights of importance to all of them; or to specify one statistic for the total weight of one, and thus zero weights for all other statistics.

Furthermore, compromises for the average can be shown to be generally feasible and worthwhile, because the allocations are insensitive to moderate changes of weights of importance (as is often true in statistics). After all, changing the relative importance by ratios of e.g., 2 or 5 should be less drastic than assigning the total weight 1 to one variable and 0 to all others, a process that implies infinite ratios of importance.

First, denote with $\Sigma_i V_{gi}^2/n_i$ the variance attainable for a statistic g with the allocations of sample sizes n_i for the i th component of variation. Then let $1 + L_g(n) = (\Sigma_i V_{gi}^2/n_i)/V_g^2(\min) = \Sigma_i C_{gi}^2/n_i$ denote the ratio of increase (with the allocation n_i) in the variance of the g th statistic over its own minimal variance, both for the same fixed Σn_i . Thus $L_g(n)$ is the *relative* loss over the minimal value of 1, and accepting the relative variances C_{gi}^2/n_i as the functions to be minimized is a critical decision; those functions seem to me more reasonable than any others that I can imagine for the functions to be combined in (1) below. For example, I prefer them to the V_{gi}^2 which depend on arbitrary units of measurement, which are removed by the $V_{gi}^2(\min)$. But in rare cases we may be faced with $V_g^2(\min) = 0$ or very small and this may make C_{gi}^2 widely large and unstable; in these

Table 3
Loss functions (1 + L) for two populations (Kish 1976)

Allocations m_i	(A)			(B)			
	(1 + L) for $W_2/W_1 = 4$			(1 + L) for 133 countries: 0.2 to 100 mm			
	$\Sigma W_i \bar{y}_i$	$\Sigma \bar{y}_i/2$	Joint	$\Sigma W_i \bar{y}_i$	$\Sigma \bar{y}_i/133$	Joint with weights	
						1:1	$I_c/I_d:1$
MW_i	1	1.56	1.28	1	6.86	3.93	
M/H	1.36	1	1.18	3.34	1	2.17	
$\alpha \sqrt{W_i}$	1.08	1.125	1.102	1.35	1.54	1.44	
$\alpha \sqrt{W_i^2 + H^{-2}}$	1.116	1.080	1.098	1.31	1.28	1.295	
$\alpha \sqrt{0.5 W_i^2 + H^{-2}}$				1.47	1.17	(1.32)	1.27
$\alpha \sqrt{2 W_i^2 + H^{-2}}$				1.20	1.44	(1.32)	1.28
$\alpha \sqrt{4 W_i^2 + H^{-2}}$				1.12	1.66	(1.39)	1.23

In (A) there are two strata and domains ($W_1 = 0.8$ and $W_2 = 0.2$); note that the allocation $m_i = \sqrt{W_i}$ does almost as well for the joint loss as the optimal.

In (B) we have the populations of 133 countries, ranging in size from 0.2 to over 100 millions, a range of 500 in relative sizes. From this problem of allocation (for the World Fertility Survey) we omitted, for practical reasons, the four largest countries and a few under 0.2 millions. Their inclusion would raise the variance of relative sizes, W_i , from 2.5 to 12, and would make the results more dramatic. Note that the $\sqrt{W_i}$ allocation reduces losses quite well. Some compromise is better than none. But the optimal allocation, $\sqrt{W_i^2 + H^{-2}}$, is considerably better. Different values of $I_c/I_d (= 1/2, 2/1$ and $4/1)$ increase slightly the variance of the joint loss function with (1:1) weights; but they remain steady for joint loss functions with their own weights $I_c/I_d:1$.

Two examples in Table 3 illustrate the surprisingly good compromises between conflicting allocations yielded by the method of weighted averaging; its results on the fourth row of Table 3 compare very favorably with the others. The reasons for the excellent results come from the very broad flat surfaces for the optimal allocations, as discussed in Section 2 and shown elsewhere (Kish 1976; Kish 1987). For example, in Canada the 10 provinces vary seventy-fold from smallest to largest population sizes, and thus resemble B in Table 3; they serve as a graphical illustration in Figure 2. (See also Fellegi and Sunter 1974.)

cases assign arbitrary values to the C_{gi}^2 or to the I_g below. These and the following including Table 3 are developed and discussed by Kish (1976).

Then with the weights I_g assigned for relative importance of the g th statistic for any set of allocations n_i of the sample sizes,

$$\begin{aligned}
 1 + L(n) &= \Sigma_g I_g (1 + L_g(n)) = \Sigma_g I_g \Sigma_i C_{gi}^2/n_i \\
 &= \Sigma_i \Sigma_g I_g C_{gi}^2/n_i = \Sigma_i Z_i^2/n_i.
 \end{aligned}
 \tag{1}$$

After changing the order of summation, we created the new variables $Z_i^2 = \Sigma_g I_g C_{gi}^2$. This function may be minimized to give compromise solutions for fixed total cost $\Sigma_i c_i n_i$. For the conflict between $n_d = n/H$ of equal sample sizes for domains versus $n_d = nW_d$ proportional to domain sizes W_d , the optimal compromise allocations are found to be proportional to $\sqrt{W_d^2 + H^{-2}}$, with equal values for I_g .

An important example was provided by the (otherwise excellent) World Fertility Surveys, which used roughly equal sample sizes for small and large countries: actual sample sizes varied only within the range of 3 to 10 thousand and with no discernible correlation with population size. Consequently, there were two- or three-fold increases of variances in the continental averages of national surveys, their "main contributions to knowledge":

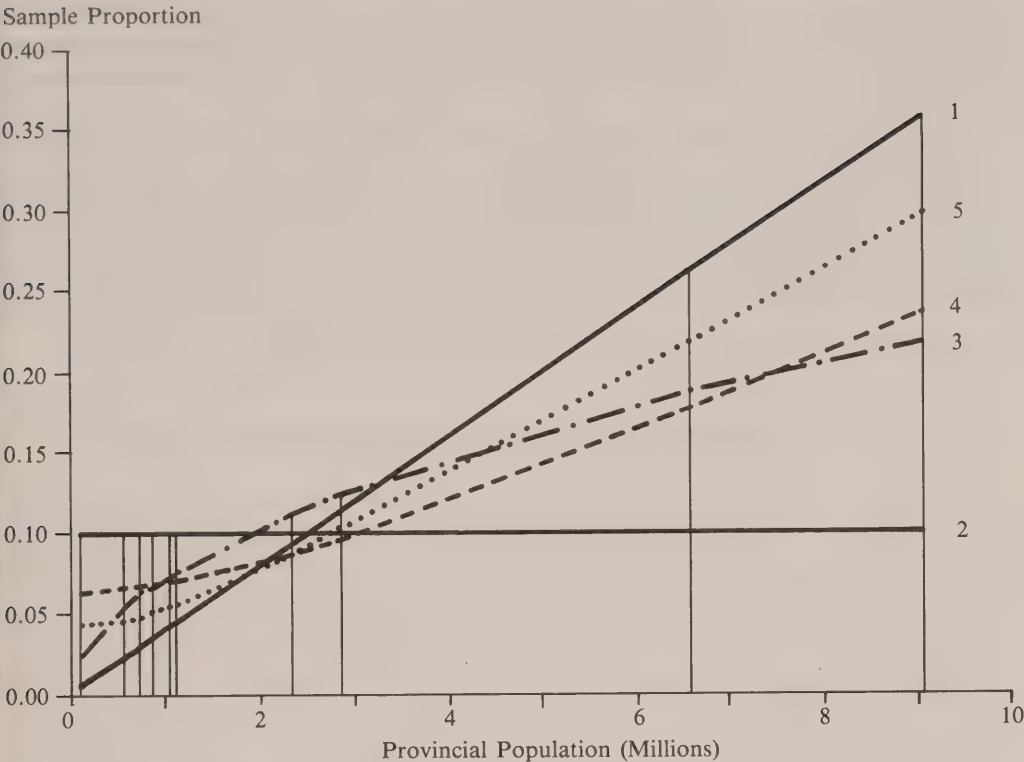


Figure 2. Five Alternative Allocations of Sample Sizes n_h of Fixed Total Σn_h

The ten provinces of Canada illustrate graphically the usual conflicts from major domains with unequal sizes, also the feasible successful compromises.

- 1 Allocation proportional to domain sizes $n_h \propto W_h$ is diagonal.
- 2 Equal allocation $n_h \propto 1/H$ is a horizontal.
Divergences of the two allocations are large near the ends.
- 3 The square-root allocation, $n_h \propto \sqrt{W_h}$ yields compromises at both ends.
- 4 The “optimal” allocation $n_h \propto \sqrt{(W^2 + 1/H^2)}$ improves both ends, and especially with an appealing “floor” near the lower end.
- 5 A “weighted” optimal $n_h \propto \sqrt{(.8W^2 + .2/H^2)}$ improves the upper end considerably.

“So far, the main contribution to knowledge has been to confirm the downward trend in fertility that characterized much of Asia and Latin America in the 1970’s and to highlight the contrast with Africa where both fertility and the desire for large numbers of children remain high” (Macura and Cleland 1985).

6. ALLOCATIONS TO STRATA AND CHOICE OF STRATIFIERS

Domains and strata often get confused in discussions, but the two aspects should be kept distinct in practical work on designs. Domains refer to subpopulations for which separate estimates are sought, whereas strata are usually smaller partitions created for decreasing variances. For example, within provinces as domains more strata may be created to reduce province variances; but cross-domains like age, sex and economic status tend to straddle across the strata. Allocations of sample sizes to strata, though often not as crucial as allocations to domains, may be important in case

of efficient disproportionate optimal allocations. The two methods of Section 5 for allocating sample sizes to domains can also be applied to allocations to strata, although the aims differ. Some of the references on nonlinear programming refer to domains and others to strata, and some confuse the two.

The presence of several survey variables and statistics among the purposes have clear implications for using more stratifying variables. Different survey variables will tend to have diverse optimal relations with the stratifiers; then it is best to use many stratifiers, even if each stratifier is used with only few stratum divisions (categories). Multipurpose design is the best reason for multivariate stratification (Kish and Anderson 1978). It may also best justify the need for "controlled selection" methods. The choice of stratum boundaries, called "optimal stratification", is a related topic, but of less importance in this condensed presentation.

7. CLUSTER SIZES; MEASURES OF SIZE; RETAINING UNITS

In descriptions of sample designs we find sometimes that the design effect has been approximated with $D_g^2 = [1 + \rho_g(\bar{b}_t - 1)]$, where ρ stands for a synthetic intraclass correlation of the "most important" variable g and $\bar{b}_t = n/a$, the average cluster size. This would yield the effective element variance $S_g^2 D_g^2$ and the variance $S_g^2 D_g^2 / n$ for the mean of the variable g . However, we must question the contents of n and of \bar{b}_t . If our population consists of married women of childbearing age, they may be only 10 percent of total persons and found in only 30 percent of dwellings; and much fewer than that for some rare populations. This situation has been treated in sampling for rare traits (Kish 1965). "Ordinarily we avoid large clusters, because of their adverse effects on the variance. But even large clusters of the entire population will yield only small clusters of a rare trait, if this is widely spread. For example, entire blocks may be sampled for persons over 65 years of age; entire villages may be searched for persons with an identifiable disease. If, on the contrary, the trait is concentrated in small areas, those areas often can be recognized and stratified accordingly."

In multipurpose designs, the crossclasses of the sample will be of variable sizes that are portions of the total sample size n_t , with \bar{M}_g as their different proportions in the populations. Thus we want to estimate in the design not only $[1 + \rho_g(\bar{b}_t - 1)]$ for diverse variables g for the total sample n_t , but also $[1 + \rho_g(\bar{b}_t - 1)]$ for many crossclasses. Here, as in Section 6, the index g is made to serve both variables and subclasses, in order to simplify notation. Then we make use of some conjectures that have been shown to be good approximations in thousands of empirical computations for scores of samples:

$$[1 + \rho_g(\bar{b}_g - 1)] \approx [1 + \rho_g(\bar{M}_g \bar{b}_t - 1)] \approx [1 + \rho_t(\bar{M}_g \bar{b}_t - 1)] \quad (2)$$

That is, we use $\bar{b}_g = \bar{M}_g \bar{b}_t$ and $\rho_g \approx \rho_t$ as rough approximations. True that this somewhat underestimates the average values of D_g^2 for crossclasses, because of variations in cluster sizes of crossclasses. But that is a small factor compared to the large variations of ρ_g between variables (Kish 1987; Verma *et al.* 1980; Kish *et al.* 1976), and that underestimate has small effects on the efficiency of designs. It is important to consider efficiencies of estimates for subclasses as well as for the entire sample; these considerations point to considerably higher efficiencies for larger clusters than would be shown for \bar{b}_t and n_t for the total sample only.

Measures of size are related to cluster sizes, but differ because of errors in the available measures, due especially to different population contents and to obsolescence. We must also note problems concerning measures of size for multisubject surveys and for "integrated survey operations" for

different populations, which may especially need drastic compromises. Those two levels of purposes (Table 1) should be distinguished because multisubject surveys use single samples in one operation; but integrated survey operations may use different sizes of sampling units for different surveys (United Nations 1980). For example, consider integrated designs for total populations and for agriculture; also perhaps for ethnic subpopulations; also perhaps for industrial or business activities: the measures of size for each of these may differ greatly. Yet some compromise solution may be found to yield reasonable efficiencies for each.

Measures of size are also closely related to problems for "Retaining units after changing strata and probabilities" (Kish and Scott 1971). Those methods were designed to deal with changes over time of sampling units, both in measures of size and in stratifying variables; but the methods are also relevant for differences between survey variables:

"Unequal selection probabilities are often assigned to sampling units. Our methods, though more generally applicable, are especially needed for the selection of primary sampling units for surveys. Often these are selected separately from many strata, with one selection from each stratum.

"After the initial selection the units may be used for many surveys over several years. But as time passes, the needs of new surveys may be better served by new strata and new selection probabilities, based on new data, than by those used for the initial selection. The difference between initial and new data may be due to differential changes among the sampling units as revealed by the latest Census. Or the differences may be due to changes in survey objectives and populations; for example, a sample initially designed for households and persons may later be required to serve a survey of farmers, or college students. *Obviously our methods are also applicable to designing simultaneously a related group of samples with differing objectives.*"

This method allows for using the best measures (for size and for strata) separately for each sample purpose, but maximizing the retention of the overlap of sampling units between the samples for separate purposes (especially PSU's). However, it would be possible to design a compromise that would average the measures in order to achieve a complete overlap of units, but sacrificing some efficiency for each of the purposes. A compromise between the two techniques may be even better than either: increase the overlap with small sacrifices of separate efficiencies by recognizing only differences of measures that surpass some arbitrary minimal criteria (Kish and Scott 1971).

8. PURPOSES AND DESIGNS FOR PERIODIC STUDIES

Periodic studies provide areas of conflict with great and growing importance as their numbers and sizes increase. It is wrong to assume that those expensive and influential surveys have only one of the five purposes listed in Table 4, because usually they are needed for several or all, if the design permits their use.

In Table 4 we note five purposes and six designs. The first four are paired with similar letters on the same four lines. These pairings call attention to designs that best serve, with reduced variances, each of the four purposes. Most periodic studies have several purposes and thus we should face, and perhaps solve, the difficult problems of multipurpose designs. Actually current levels (A) and net changes (C) can be served with any of the six listed designs, but with some increase in variances or in costs. However, individual (gross, micro) changes (D) need panels; and cumulations (B) need some changes of samples, and are fastest without any overlaps. For current levels (A) variances can be somewhat reduced with estimators using correlations from partial overlaps. Net changes (C) benefit from correlations from any overlap, and most from complete overlaps (Cochran 1977; Kish 1987; Kish 1965). Reasonable compromises often become possible, when purposes can be defined. However, extraneous considerations may rule out some designs (e.g., overlaps may be either prohibited or enforced) and thus force the use of less efficient — but still valid — designs.

Table 4
Purposes and Designs for Periodic Samples

Purposes	Designs	Rotation Scheme
A. Current levels	A. Partial overlaps $0 < P < 1$	abc-cde-efg
B. Cumulations	B. Nonoverlaps $P = 0$	aaa-bbb-ccc
C. Net changes (means)	C. Complete overlaps $P = 1$	aaa-aaa-aaa
D. Gross changes (individual)	D. Panels	same elements
E. Multipurpose time series	E. Combinations, SPD	
	F. Master Frames	

The chief variation in these six designs concerns the amount (and kind) of overlaps between periods. The rotation scheme of complete overlaps shows, with aaa-aaa, that the periods have all common parts; the nonoverlap with aaa-bbb shows none; and the partial overlap abc-cde-efg shows c and e as 1/3 overlaps between succeeding periods only. This section concentrates on the effects of varying proportions of overlaps P in diverse designs on different purposes; in complete overlaps $P = 1$, in nonoverlaps $P = 0$, and in partial overlaps $0 < P < 1$. The purposes are discussed in terms of variances for estimated means, because means (and percentages, rates, proportions) are both the most used and the simplest estimates. Effects on other estimates will not be entirely different but they are too many, diverse, and difficult to be explored here.

More discussions of panels is also available elsewhere, with its advantages, disadvantages, problems and solutions (Duncan and Kalton 1986; Kish 1987). I call attention to SPD, or Split Panel Designs, that I am trying to promote for multipurpose designs. These would combine a panel sample P with new rotating or “rolling” samples, so that $Pa-Pb-Pc-Pd$ would symbolize the periodic samples. The rolling samples a,b,c,d etc., could be cumulated into larger samples. The panel P serves primarily to provide micro (individual gross changes). But it also serves as the partial overlap for better estimates of both current levels and macro (mean, net) changes for any pair of periods.

9. COMPUTING AND PRESENTING SAMPLING ERRORS

It seems questionable to include this topic under design, but I have no doubt that this is a multipurpose problem. The strategies for computing and presenting sampling errors deserve separate listing as an area of conflict among the many statistics given generally for the results of surveys. It is not enough to present standard errors for only one or a few of the most important statistics: they are too many and too diverse. Because of that diversity, the practice has grown up to compute from the variances other expressions of sampling variability, especially estimates of the “design effects” d_g^2 ; also sometimes from the $d_g^2 = 1 + \rho_g(\bar{d}_g - 1)$, estimates of the synthetic intraclass correlation ρ_g .

Briefly, I advise: a) Compute sampling errors for many variables, because the variances, the design effects (d_g^2), and the intraclass coefficients (ρ_g) can and do differ greatly between variables. b) You may have to do some averaging of sampling errors, because it may be inconvenient or confusing to present them all. c) It may be neither feasible nor necessary to compute sampling errors for all subclasses, because they can often be approximated with reasonable models. d) It is necessary to present sampling errors for subclasses and for other statistics to guide the readers of the reports (Kish 1965; Kish 1987; Verma *et al* 1980). I hope that this topic will receive in the future from theorists and methodologists some of the attention it needs.

10. CONCLUSIONS

For the ten areas of conflict of Section 3 approaches and solutions are proposed in Sections 4 to 9 that are very diverse. Averaging allocations among domains in Section 5 seems to give surprisingly good compromise solutions. The advice in Section 6 to use more stratifiers can also yield worthwhile gains. In Sections 4 and 7 considerations for subclass estimates lead to drastically different decisions for sample designs. In Section 8 we note how periodic designs can be best suited to purposes, and best compromise for multipurpose aims. We looked at the different levels of purposes and at the various areas of conflicts jointly. Asking the right question is the core of most problems. I propose multipurpose design as a new paradigm, to replace "optimal" solutions to artificially partial questions such as: What is the optimal allocation for the mean \bar{y} or the total \bar{Y} of "the most important" variable?

REFERENCES

- BEAN, J.A., and BURMEISTER, L.F. (1978). A review of optimal sample allocation for multipurpose surveys, *Biometrika*, 20, 3-14.
- CHATTERJEE, S. (1967). A note on optimum stratification, *Skandinavisk Actuarietidskrift*, 50, 40-44.
- COCHRAN, W.G. (1977). *Sampling Techniques*, (3rd ed). New York: John Wiley and Sons.
- DALENIUS, T. (1957). *Sampling in Sweden*, Stockholm: Almquist and Wicksell.
- DUNCAN, G.J., and KALTON, G. (1986). Issues of design and analysis of surveys across time. *International Statistical Review*, 54.
- FELLEGI, I.P., and SUNTER, A.B. (1974). Balance between different sources of survey errors. *Sankhyā*, 36, 119-142.
- FOREMAN, E.K. (1983). Integrated programmes of household surveys: design aspects. *Bulletin of the International Statistical Institute*.
- KING, A.J., and JESSEN, R.J. (1945). The master sample of agriculture, *Journal of the American Statistical Association*, 38-56.
- KISH, L. (1987). *Statistical Design for Research*: New York: Wiley-Interscience.
- KISH, L. (1986). Timing of surveys for public policy. *Australian Journal of Statistics*, 28, 1-12.
- KISH, L. (1980). Design and estimation for domains. *The Statistician*, London, 29, 209-222.
- KISH, L. (1976). Optima and proxima in linear sample designs. *Journal of the Royal Statistical Society*, A, 139, 80-95.
- KISH, L. (1965). *Survey Sampling*. New York: John Wiley and Sons.
- KISH, L. (1961). Efficient allocation for multipurpose samples. *Econometrica*, 29, 363-385.
- KISH, L., and ANDERSON, D.W. (1978). Multivariate and multipurpose stratification, *Journal of the American Statistical Association*, 73, 24-34.
- KISH, L., and FRANKEL, M.R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society*, B, 36, 1-37.
- KISH, L., and SCOTT, A.J. (1971). Retaining units after changing strata and probabilities, *Journal of the Royal Statistical Society*, 66, 461-470.
- KIREGYERA, B., and GACHUKI, P. (1985). Experiences in panel surveys: examples from an integrated sample survey programme in Kenya. *Bulletin of the International Statistical Institute*.
- MACURA, M., and CLELAND, J. (1985). *A Celebration of Statistics: the ISI Centenary Volume*, (Eds. A.C. Atkinson and S.E. Fienberg), New York: Springer Verlag.

- MURTHY, M.N. (1974). Evaluation of multi-subject sample survey systems. *International Statistical Review*, 42.
- MURTHY, M.N. (1967). *Sampling Theory and Methods*. Calcutta: Statistical Publishing Society.
- UNITED NATIONS (1980). *National Household Survey Capability Programme* (NHSCP). New York: United Nations.
- RODRIQUEZ-VERA, A. (1982). *Multipurpose Optimal Sample Allocation Using Mathematical Programming*. Ph.D. dissertation, The University of Michigan, Ann Arbor.
- VERMA, V., SCOTT, C., and O'MUIRCHEARTAIGH, C. (1980). Sample designs and sampling errors for the World Fertility Survey. *Journal of the Royal Statistical Society, A*, 143, 431-473.
- YATES, F. (1981). *Sampling Methods for Censuses and Surveys*, (4th ed.). London: Griffin and Co.

On the Stratification of Skewed Populations

PIERRE LAVALLÉE and MICHEL A. HIDIROGLOU¹

ABSTRACT

For a given level of precision, Hidirolou (1986) provided an algorithm for dividing the population into a take-all stratum and a take-some stratum so as to minimize the overall sample size assuming simple random sampling without replacement in the take-some stratum. Sethi (1963) provided an algorithm for optimum stratification of the population into a number of take-some strata. For the stratification of a highly skewed population, this article presents an iterative algorithm which has as objective the determination of stratification boundaries which split the population into a take-all stratum and a number of take-some strata. These boundaries are computed so as to minimize the resulting sample size given a level of relative precision, simple random sampling without replacement from the take-some strata and use of a power allocation among the take-some strata. The resulting algorithm is a combination of the procedures of Hidirolou (1986) and Sethi (1963).

KEY WORDS: Iterative algorithm; Optimum boundaries; Take-all; Take-some.

1. INTRODUCTION

Efficient sampling of highly skewed populations such as those displayed by business surveys require that they be stratified into a take-all stratum and a number of take-some strata. The whole of units the take-all stratum is selected with certainty whereas units in the take-some strata are selected by a probability mechanism. Approximate cut-off rules for stratifying a population into a take-all and a single take-some stratum have been given by Glasser (1962) and Hidirolou (1986). Glasser (1962) provided the cut-off value under the assumption that a fixed total sample size was to be drawn from the take-all and take-some stratum, and that the take-some sampled units were to be selected without replacement using simple random sampling. Hidirolou (1986) provided the cut-off value under the assumption that a required level of precision had to be satisfied. These two approaches are dual in the sense that Glasser's objective was to minimize sampling variance for fixed sample size, whereas Hidirolou's objective was to minimize sample size for fixed sampling variance.

In this article, an algorithm for stratifying a highly skewed population into a take-all stratum and a number of take-some strata will be presented. The objective will be to minimize the overall sample size given the coefficient of variation of the estimator and the allocation scheme of the sample to the take-some strata. The strata boundaries will be derived in term of an auxiliary variable which is closely related to the information being collected by the survey. For example, for a census of retailers, if yearly sales is one of the variables measured, this auxiliary variable can be used to determine the strata boundaries for a single-purpose survey which collect sales on a monthly basis. For a multi-purpose survey, given that the strata boundaries have been determined using

¹ Pierre Lavallée is Methodologist and Michel A. Hidirolou is Chief, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario K1A 0T6, Canada. The authors would like to acknowledge France Bilocq, Business Survey Methods Division, Statistics Canada, for programming the examples.

an auxiliary variable closely related to the main variable, the optimality of these boundaries will diminish for other variables which are not well correlated with it. The algorithm is a modification of Sethi's (1963) method for stratifying a population. The resulting boundaries, which are optimal, will provide the required minimum sample size.

The allocation scheme which has been chosen to illustrate the method is the power allocation. The use of this type of allocation enables the publication of strata estimates which do not have markedly different coefficients of variation. Power allocation has been proposed by Carroll (1970), Fellegi (1981) and Bankier (1988). It is found to offer in practice a compromise between Neyman allocation and the requirement to have equal coefficients of variation for each stratum. A disadvantage of Neyman allocation is that if estimates are required for each stratum, the associated coefficients of variation may be quite different between the strata. Alternatively, an allocation which achieves equal coefficients of variation amongst the strata may require sample size which is much larger than the one required under Neyman allocation. In our context, power allocation would enable the publication of estimates for strata of varying sizes (small, medium and large) companies with similar coefficients of variation.

The method developed in the paper will be numerically compared, in terms of boundary values and sample size, to the Dalenius — Hodges (1959) cumulative square root f rule, as well as to a mixture of the Hidirolou (1986) and the Dalenius — Hodges (1959) stratification methods. The algorithm, which is recursive in nature, is simple to program and converges rapidly to the optimum boundary points. It also offers substantial savings in terms of sample size for given reliability criteria.

2. THE PROBLEM

Consider a finite ordered population of N units:

$$y_{(1)}, y_{(2)}, \dots, y_{(N)},$$

with $y_{(i)} \leq y_{(i+1)}$ for $i = 1, 2, \dots, N-1$. This population is to be stratified into L strata. The number of units in each stratum is denoted by N_h , $h = 1, 2, \dots, L$. The sampling scheme calls for n_h units to be drawn from each corresponding take-some stratum of size N_h ($h = 1, 2, \dots, L-1$) without replacement, using simple random sampling, with $n_L = N_L$. The mean to be estimated is

$$\bar{Y} = \sum_{h=1}^L \sum_{j=M_{h-1}+1}^{M_h} y_{(j)} / N, \quad (2.1)$$

where $M_h = \sum_{i=1}^h N_i$ for $h = 1, 2, \dots, L$ and $M_0 = 0$.

Given this set up, the estimator of population mean \bar{Y} is

$$\hat{\bar{Y}} = \left[\sum_{h=1}^{L-1} \frac{N_h}{n_h} \sum_{j=M_{h-1}+1}^{m_n} z_j + \sum_{j=M_{L-1}+1}^N y_{(j)} \right] / N \quad (2.2)$$

where $y_{M_{h-1}+1} \leq z_j \leq y_{M_h}$ for $j = m_{h-1}+1, \dots, m_h$ ($h = 1, 2, \dots, L-1$), $m_h = \sum_{i=1}^h n_i$ for $h = 1, 2, \dots, L$ and $m_0 = 0$.

Assume that the desired level of precision for the estimated mean is specified by c (coefficient of variation) and that the proportion of sampled units to be allocated to each of the first $L-1$ strata is a_h ($h = 1, 2, \dots, L-1$) where $\sum_{h=1}^{L-1} a_h = 1$. The term " a_h " is conveniently used to represent any type of allocation to the strata. For instance, in the case of N -proportional power allocation,

$$a_h = \frac{N_h^p}{\sum_{h=1}^{L-1} N_h^p} \quad (h = 1, 2, \dots, L-1)$$

and in the case of Y -proportional power allocation,

$$a_h = \frac{Y_h^p}{\sum_{h=1}^{L-1} Y_h^p},$$

where $0 < p < \infty$. The power allocations have the property that under relatively simple assumptions and for a suitable choice of p , the coefficients of variation for the take-some strata tend to be equalized without a significant increase in the overall coefficient of variation. This equality of coefficients of variation is often desired by the users of the survey data.

In practice, the value of p is often chosen to be $1/2$ or $1/3$. A small value of p (i.e. p close to 0) usually yields similar stratum coefficients of variation while a larger value increases the discrepancy between the coefficients of variation but also increases the precision of the overall estimates.

It would be noted that these power allocations are equivalent to the allocation proposed by Bankier (1988) when the population coefficients of variation of the take-some strata are equal.

The variance of \hat{Y} is

$$V(\hat{Y}) = \frac{1}{N^2} \sum_{h=1}^{L-1} \frac{N_h}{n_h} (N_h - n_h) S_h^2, \quad (2.3)$$

where S_h^2 denotes the population variance of each stratum h . In terms of the desired level of coefficient of variation c , $V(\hat{Y})$ may be reexpressed as $V(\hat{Y}) = c^2 \bar{Y}^2$. Substituting $n_h = (n - N_L) a_h$ and $V(\hat{Y}) = c^2 \bar{Y}^2$ into (2.3) and solving for n obtains

$$n = N_L + \frac{\sum_{h=1}^{L-1} N_h^2 S_h^2 / a_h}{(N c \bar{Y})^2 + \sum_{h=1}^{L-1} N_h S_h^2}. \quad (2.4)$$

The problem is to find boundaries $b_{(1)}, b_{(2)}, \dots, b_{(L-1)}$ (where $y_{(1)} < b_{(1)} < \dots < b_{(L-1)} < y_{(N)}$) such that the overall sample size n is minimized, given the level of reliability c and the specific allocation scheme (represented by a_h).

3. THE ALGORITHM

The approach used in this paper, for obtaining stratification boundaries for a desired level of precision, has first been used by Dalenius (1950) in the case of stratification boundaries for a given sample size. It is first assumed that the sampling is done from a population whose frequency distribution may with sufficient accuracy be represented by a continuous density $f(y)$. Then, for a given set of boundaries $b_{(1)}, \dots, b_{(L-1)}$ the following quantities are defined:

$$W_h = \int_{b_{(h-1)}}^{b_{(h)}} f(y) \, dy, \tag{3.1}$$

$$\mu_h = \int_{b_{(h-1)}}^{b_{(h)}} y f(y) \, dy / W_h, \tag{3.2}$$

$$\sigma_h^2 = \int_{b_{(h-1)}}^{b_{(h)}} y^2 f(y) \, dy / W_h - \mu_h^2, \tag{3.3}$$

for $h = 1, \dots, L$, with $b_{(0)} = -\infty, b_{(L)} = +\infty$.

Equation (2.4) can then be rewritten as

$$n = NW_L + \frac{N \left(\sum_{h=1}^{L-1} W_h^2 \sigma_h^2 / a_h \right)}{N c^2 \mu^2 + \sum_{h=1}^{L-1} W_h \sigma_h^2}, \tag{3.4}$$

where

$$\mu = \int_{b_{(0)}}^{b_{(L)}} y f(y) \, dy.$$

It should be noted that even if the population is considered to be large, the finite population correction (f.p.c.) factor is still present in equation (3.4) - see Dalenius-Gurney (1951). By definition, the take-all stratum needs to have a finite population in order to get a finite sample size. Also, ignoring the f.p.c. would not lead to a zero variance for the take-all stratum.

The a_h in equation (2.3) can also be represented using the quantities (3.1), (3.2) and (3.3). In the case of the N -proportional power allocation, we get:

$$a_h = \frac{W_h^p}{\sum_{h=1}^{L-1} W_h^p}, \tag{3.5}$$

for $h = 1, \dots, L-1$.

For the Y -proportional power allocation, the following is obtained:

$$a_h = \frac{(W_h \mu_h)^p}{\sum_{h=1}^{L-1} (W_h \mu_h)^p}, \quad (3.6)$$

where $0 < p < \infty$.

In this paper, the Y -proportional power allocation will mainly be considered but the calculations can also be performed for the N -proportional power allocation and, in fact, for any kind

of allocation represented by some a_h where $\sum_{h=1}^{L-1} a_h = 1$. Putting equation (3.6) into (3.4), we get

$$n = N W_L + \frac{N \left[\sum_{h=1}^{L-1} (W_h \sigma_h)^2 (W_h \mu_h)^{-p} \right] \left[\sum_{h=1}^{L-1} (W_h \mu_h)^p \right]}{N c^2 \mu^2 + \sum_{h=1}^{L-1} W_h \sigma_h^2}. \quad (3.7)$$

In order to find the optimal boundaries $b_{(1)}, \dots, b_{(L-1)}$ such that the sample size n will be minimum, the derivatives of equation (3.7) are taken with respect to $b_{(1)}, \dots, b_{(L-1)}$, respectively, and equated to zero. The resulting equations are:

For $h = 1, \dots, L-2$,

$$\begin{aligned} & [F T_h - F T_{h+1}] b_{(h)}^2 + \\ & [F K_h - 2\mu_h F T_h - F K_{h+1} + 2\mu_{h+1} F T_{h+1} + 2\mu_h AB - 2\mu_{h+1} AB] b_{(h)} + \\ & [F T_h \mu_h^2 + F T_h \sigma_h^2 - F T_{h+1} \mu_{h+1}^2 - F T_{h+1} \sigma_{h+1}^2 - AB\mu_h^2 + AB\mu_{h+1}^2] = 0, \end{aligned} \quad (3.8)$$

and for $h = L-1$,

$$\begin{aligned} & [F T_{L-1} - AB] b_{(L-1)}^2 + \\ & [F K_{L-1} - 2\mu_{L-1} F T_{L-1} + 2\mu_{L-1} AB] b_{(L-1)} + \\ & [F T_{L-1} \mu_{L-1}^2 + F T_{L-1} \sigma_{L-1}^2 - AB\mu_{L-1}^2 - F^2] = 0, \end{aligned} \quad (3.9)$$

where

$$A = \sum_{h=1}^{L-1} (W_h \mu_h)^p,$$

$$B = \sum_{h=1}^{L-1} (W_h \sigma_h)^2 (W_h \mu_h)^{-p},$$

$$F = N c^2 \mu^2 + \sum_{h=1}^{L-1} W_h \sigma_h^2,$$

$$K_h = B p (W_h \mu_h)^{p-1} - A p (W_h \sigma_h)^2 (W_h \mu_h)^{-p-1},$$

$$T_h = A W_h (W_h \mu_h)^{-p}.$$

Labeling the coefficient of $b_{(h)}^2$ as α_h , the coefficient of $b_{(h)}$ as β_h and the remaining terms as γ_h , equations (3.8) and (3.9) can be represented as quadratic equations of the form $\alpha_h b_{(h)}^2 + \beta_h b_{(h)} + \gamma_h = 0$. However, as pointed out by Sethi (1963), the terms α_h , β_h and γ_h are themselves functions of $b_{(1)}$, . . . , $b_{(L-1)}$ through the integrals (3.1), (3.2) and (3.3). Using Sethi's (1963) approach, equations (3.8) and (3.9) can easily be solved using the following iterative method:

STEP 1 : Start with some arbitrary boundaries $b'_{(1)} < \dots < b'_{(L-1)}$.

STEP 2 : Calculate the proportions W'_h , the means μ'_h and the variances $\sigma_h'^2$ (from equations (3.1), (3.2) and (3.3), respectively) based on these boundaries, $h = 1, \dots, L-1$.

STEP 3 : Replace the initial set of boundaries by $b''_{(1)}, \dots, b''_{(L-1)}$ where

$$b''_{(h)} = \frac{-\alpha'_h + \sqrt{\beta_h'^2 - 4\alpha'_h\gamma'_h}}{2\alpha'_h}, \quad h = 1, \dots, L-1. \quad (3.10)$$

STEP 4 : Repeat steps 2 and 3 till two consecutive sets are either identical or differ by negligible quantities, i.e.

$$\max_{h=1}^{L-1} |b''_{(h)} - b'_{(h)}| < \epsilon \text{ for some } \epsilon > 0. \quad (3.11)$$

It should be noted that it can be proved that the sign before the square root ($\sqrt{\quad}$) is positive because $b'_{(h)}$ lies between μ'_h and μ'_{h+1} .

The difficulty of using the above algorithm is that some knowledge of $f_{(y)}$, the approximate density, is required. Since the population considered is finite, it is possible to overcome this difficulty by replacing the quantities (3.1), (3.2) and (3.3) by corresponding expressions based on the finite population. Hence, proceeding as in Cochran (1977), the infinite population parameters given by expressions (3.1), (3.2) and (3.3) can be replaced by their finite population counterparts. That is:

$$W_h = \frac{N_h}{N}, \quad (3.12)$$

$$\bar{Y}_h = \frac{1}{N_h} \sum_{j=b_{(h-1)+1}}^{b_{(h)}} y_{(j)}, \quad (3.13)$$

$$S_h^2 = \frac{1}{N_{h-1}} \sum_{j=b_{(h-1)+1}}^{b_{(h)}} y_{(j)}^2 - N_h \bar{Y}_h^2, \quad (3.14)$$

for $h = 1, \dots, L$.

Using these last quantities, the problem described in section 2 of finding boundaries $b_{(1)}, \dots, b_{(L-1)}$ such that the overall sample size n is minimized for a given level of reliability c and a specific allocation scheme can easily be solved by the following iterative method:

STEP 0 : Sort the population y_1, \dots, y_N in ascending order and set $b_{(0)} = y_{(1)}$ and $b_{(L)} = y_{(N)}$.

STEP 1 : Start with some arbitrary boundaries such that $b_{(0)} < b'_{(1)} < \dots < b'_{(L-1)} < b_{(L)}$.

STEP 2 : Calculate the proportions W'_h , the mean \bar{Y}'_h and the variance $S_h^{2'}$ (from equations (3.12), (3.13) and (3.14) respectively) based on these boundaries, $h = 1, \dots, L-1$.

STEP 3 : Replace the initial set of boundaries by $b''_{(1)}, \dots, b''_{(L-1)}$ where

$$b''_{(h)} = \frac{-\alpha'_h + \sqrt{\beta_h'^2 - 4\alpha'_h\gamma_h}}{2\alpha'_h}, \quad h = 1, \dots, L-1.$$

STEP 4: Repeat step 2 and 3 till two consecutive sets are either identical or differ by negligible quantities, i.e.

$$\max_{h=1}^{L-1} |b''_{(h)} - b'_{(h)}| < \epsilon \text{ for some } \epsilon < 0.$$

The use of this algorithm with real data will be compared to others in the next section.

4. SOME ILLUSTRATIONS

In order to display results given in Section 3, we will use data obtained from the Annual Retail Trade and Wholesale Trade Surveys conducted at Statistics Canada. These surveys measure the sales of companies whose principal business is retailing or wholesaling respectively. Three populations have been used to illustrate the algorithm. They are, respectively, other products in Wholesale in Quebec (Population 1), other foods in Wholesale in Manitoba (Population 2), and appliances, television, radio and stereo stores in Retail in Quebec (Population 3). Those populations have been chosen to reflect different combinations of population sizes: high, medium and low. The skewness for these populations is 24.2 (for Population 1), 6.5 (for Population 2) and 13.6 (for Population 3).

The numerical results provided by the algorithm will be compared to those obtained using two other methods. The first method is to simply stratify the population using the cumulative square root f rule given by Dalenius-Hodges (1959). The second method is to determine the cut-off boundary between take-all and take-some strata using the approximation given by Hidirolou (1986)

and then to apply the cumulative square root f rule to stratify the non take-all population into a number of take-some strata. The different methods will respectively be labelled as i) Cum $f^{1/2}$ rule, ii) mixture, and iii) optimum, for the currently proposed algorithm. The sole use of the Dalenius-Hodges (1959) method is not realistic because it would, in practice, only be used after the take-all stratum had been identified using some given arbitrary rule. However, we display the sole use of this method to caution against its blind use in the context of highly skewed populations.

The Hidiroglou (1986) cut-off point is obtained via the following iterative process:

$$b_{TA}'' = \mu_{[N-t']} = \left\{ \frac{N-t'-1}{(N-t')^2} N^2 c^2 \bar{Y}^2 + S_{[N-t']}^2 \right\}^{1/2},$$

(4.1)

where

$$\mu_{[N-t']} = \frac{1}{N-t'} \sum_{i=1}^{N-t'} y_{(i)}$$

(4.2)

Table 1
Effect of Varying Coefficient of Variation and Power Allocation
on Sample Sizes for Three Stratification Methods
(Population 1 — Size = 1221)

			Stratification Method								
<i>c</i>	<i>p</i>	Strata	Cum $f^{1/2}$ Rule			Mixture			Optimum		
			N_h	n_h	$b_{(h)}$	N_h	n_h	$b_{(h)}$	N_h	n_h	$b_{(h)}$
0.05	0.25	1	1196	177*		1017	16		891	11	
		2	20	20	3,715,320	152	14	465,180	290	13	302,912
		3	5	5	14,786,280	52	52	1,131,961	40	40	1,835,930
		Total		202			82			64	
0.05	0.50	1	1196	178*		1017	16		863	10	
		2	20	20	3,715,320	152	13	465,180	318	14	289,422
		3	5	5	17,786,280	52	52	1,131,961	40	40	1,832,038
		Total		203			81			64	
0.01	1.00	1	1196	616*		751	37		687	36	
		2	20	20	3,715,320	215	34	196,840	374	78	162,068
		3	5	5	14,786,280	255	255	383,033	160	160	564,076
		Total		641			326			274	
0.05	1.00	1	1196	180*	3,715,320	1017	16		858	8	
		2	20	20	14,786,280	152	11	465,180	323	16	271,920
		3	5	5		52	52	1,131,961	40	40	1,867,254
		Total		205			79			64	
0.10	1.00	1	1196	56*		1073	7		1007	7	
		2	20	20	3,715,320	109	4	592,900	191	9	442,357
		3	5	5	14,786,280	39	39	1,953,113	23	23	4,032,950
		Total		81			50			39	

*Requires over allocation to satisfy coefficient of variation.

Table 2
Effect of Varying Coefficient of Variation and Power Allocation
on Sample Sizes for Three Stratification Methods
(Population 2 — Size = 44)

c	p	Strata	Stratification Method								
			Cum $f^{1/2}$ Rule			Mixture			Optimum		
			N_h	n_h	$b_{(h)}$	N_h	n_h	$b_{(h)}$	N_h	n_h	$b_{(h)}$
0.05	0.25	1	42	38		32	1		29	1	
		2	1	1*	137,939,900	6	1	4,708,409	11	1	3,029,455
		3	1	1	459,739,000	6	6	10,622,301	4	4	17,461,464
		Total		40			8			6	
0.05	0.50	1	42	38		32	1		28	1	
		2	1	1*	137,939,900	6	1	4,708,409	12	1	2,582,819
		3	1	1	459,739,000	6	6	10,622,301	4	4	17,640,325
		Total		40			8			6	
0.01	1.00	1	42	42		25	1		25	1	
		2	1	1	137,939,900	5	1	1,059,550	10	4	1,153,322
		3	1	1	459,739,000	14	14	3,742,377	9	9	5,969,271
		Total		44			16			14	
0.05	1.00	1	42	38		32	1		26	1	
		2	1	1*	137,939,900	6	1	4,708,409	14	2	1,779,500
		3	1	1	459,739,000	6	6	10,622,301	4	4	17,349,902
		Total		40			8			7	
0.10	1.00	1	42	30		34	1		28	1	
		2	1	1*	137,939,900	6	1	4,848,218	13	1	2,413,800
		3	1	1	459,739,000	4	4	16,749,625	3	3	30,091,449
		Total		32			6			5	

*Requires over allocation to satisfy coefficient of variation.

and

$$S^2_{[N-t']} = \frac{1}{N-t'-1} \sum_{i=1}^{N-t'} (y_{(i)} - \mu_{[N-t']})^2.$$

The number of take-all units obtained for each step of this iterative process is t' . The starting point for this approximation is

$$b'_{TA} = \mu_{[N]} + \{Nc^2 \bar{Y}^2 + S^2_{[N]}\}^{1/2} \tag{4.3}$$

The stopping point for (4.1) is reached when the following inequality is satisfied:

$$0 \leq 1 - n(t'')/n(t') < 0.10 \tag{4.4}$$

Table 3
Effect of Increasing the Number of Strata on
Sample Sizes for Two Stratification Methods
 $p = 1, c = 0.05$

Population 1 ($N = 1221$) Stratification Method		Number of Strata								
		3			4			5		
	Strata	N_h	n_h	$b_{(h)}$	N_h	n_h	$b_{(h)}$	N_h	n_h	$b_{(h)}$
Mixture	1	1017	16		897	6		823	3	
	2	152	11	465,180	194	5	311,117	194	2	245,090
	3	52	52	1,131,961	78	4	641,252	101	2	465,180
	4				52	52	1,131,961	51	2	751,297
	5							52	52	1,131,961
	Total		79			67			61	
Optimum	1	858	8		704	3		655	2	
	2	323	16	271,920	373	7	173,981	358	4	149,327
	3	40	40	1,867,254	112	6	604,869	163	5	453,114
	4				32	32	2,676,449	29	4	1,522,329
	5							16	16	5,810,487
	Total		64			48			31	

Population 3 ($N = 161$)										
Mixture	1	106	6		84	2		71	1	
	2	39	6	265,480	38	2	185,320	35	1	155,260
	3	16	16	553,255	23	2	335,620	22	1	265,480
	4				16	16	553,255	17	1	385,720
	5							16	16	553,255
	Total		28			22			20	
Optimum	1	86	4		55	1		34	1	
	2	65	9	199,415	61	3	125,572	51	1	83,594
	3	10	10	680,942	39	5	312,769	42	2	192,215
	4				6	6	826,942	29	3	382,236
	5							5	5	906,894
	Total		23			15			12	

where

$$n(t') = t' + \frac{(N-t')^2 S^2_{[N-t']}}{(Nc\bar{Y})^2 + (N-t') S^2_{[N-t']}} \quad (4.5)$$

Tables 1 and 2 display the results for a large population (Population 1) and a small population (Population 2) for a number of different coefficients of variation and power allocations. Table 3 displays the results for the large population (Population 1) and a medium population (Population 3) by varying the number of strata. For all three tables, the allocation of the sample to the take-some strata is the power Y -proportional scheme.

The following conclusions can be drawn from Tables 1 and 2. The use of the cumulative square root f rule to determine boundary points is very inefficient in the present context. Substantial gains,

in terms of sample size reduction, are made by using the mixture rule. For the three strata used in those two tables, further reductions in sample size of the order of 20% can be achieved by using the optimum rule. For a given fixed coefficient of variation, the variation of the power " p " has a minor impact on the resulting sample size. As expected, sample sizes increase when the required coefficient of variation, c , is decreased (for a fixed power allocation). The optimum method declares less take-all units (stratum 3) than the mixture method, or stated another way, the take-all stratum boundary is higher for the optimum than for the mixture. The cumulative square root rule loses its efficiency in the take-all stratum boundary determination. It is readily observed that the boundary for this method is significantly higher than those obtained with the other methods.

In Table 3, we only compare the mixture and optimum methods for two populations, varying the number of strata, for a fixed coefficient of variation and Y -proportional power allocation. Similar conclusions to those drawn from Tables 1 and 2 hold. The effect of increasing the number of strata is to reduce the number of sampled units for both methods. However, the reduction becomes more pronounced for the optimum method as the number of strata increases.

5. CONCLUSION

The optimal stratification, of a skewed population into a take-all stratum and a number of take-some strata, has provided a substantial reduction in overall sample size for given relative precision. The method can be adapted to any type of allocation and to any number of strata. The take-all condition can also be excluded.

The algorithm, which is recursive in nature, converges quickly. It is simple to implement on the computer using SAS, FORTRAN, or any other high level language.

REFERENCES

- BANKIER, M.D. (1988). Power allocations, determining sample sizes for sub-national areas. To appear in *The American Statistician*.
- CARROL, J. (1970). Allocation of a sample between States. Unpublished memorandum of Australian Bureau of Census and Statistics.
- COCHRAN, W.G. (1977). *Sampling Techniques*, (3rd. ed.). New York: John Wiley & Sons.
- DALENIUS, T. (1950). The problem of optimum stratification. *Skandinavisk Aktuarietidskrift*, 33, 203-213.
- DALENIUS, T., and GURNEY, M. (1951). The problem of optimum stratification. *Skandinavisk Aktuarietidskrift*, 34, 133-148.
- DALENIUS, T., and HODGES, J.L.Jr. (1959). Minimum variance stratification, *Skandinavisk Aktuarietidskrift*, 54, 88-101.
- FELLEGI, I.P. (1981). Should the census counts be adjusted for allocation purposes? - Equity considerations. In *Current Topics in Survey Sampling* (Eds. D. Krewski, R. Platek and J.N.K. Rao). New York: Academic Press, 47-76.
- GLASSER, G.J. (1962). On the complete coverage of large units in a statistical study. *International Statistical Review*, 30, 28-32.
- HIDIROGLOU, M.A. (1986). The construction of a self-representing stratum of large units in survey design. *The American Statistician*, 40, 27-31.
- SETHI, V.K. (1963). A note on optimum stratification of populations for estimating the population means. *Australian Journal of Statistics*, 5, 20-33.

Research Issues in the Survey of Income and Program Participation¹

DANIEL KASPRZYK²

ABSTRACT

The Survey of Income and Program Participation (SIPP) is an ongoing nationally representative household survey program of the Bureau of the Census. The primary purpose of the SIPP is to improve the measurement of information related to the economic situation of households and persons in the United States. It accomplishes this goal through repeated interviews of sample individuals using a short reference period and a probing questionnaire. The multi-interview design of the SIPP raises methodological and statistical issues of concern to all panel surveys of families and persons. This paper reviews these issues as they relate to the SIPP. The topics reviewed are: 1) questionnaire design; 2) data collection, including respondent rules, data collection mode, length of reference period, and rules for following movers; 3) concepts, design, and estimation; and 4) response error.

KEY WORDS: Panel surveys; Questionnaire design; Survey design; Longitudinal estimation; Response error.

1. INTRODUCTION

The Survey of Income and Program Participation (SIPP) is an ongoing nationally representative household survey program of the U.S. Bureau of the Census. It provides comprehensive information on the economic resources of the American people and on how public transfer and tax programs affect their financial circumstances. The data from the SIPP provide government policy makers with an information base for studying the efficiency of government tax and transfer programs, for estimating future program costs and coverage, and for assessing the effects of proposed policy changes. The SIPP is designed to improve the measurement of information related to the economic situation of households and persons in the United States, and is the culmination of a large-scale development program, the Income Survey Development Program (ISDP), which examined concepts, procedures, questionnaires, recall periods, and the like (Ycas and Lininger, 1981).

The need for a survey like SIPP arose because of the limitations of the March Income Supplement of the Current Population Survey (CPS), the principal source of information on the distribution of household and personal income in the United States. These limitations are inherent in the survey design, survey instrument, and survey procedures and can not be easily modified. As a consequence the Income Survey Development Program was established in 1975 by the U.S. Department of Health and Human Services to develop methods to overcome the principal shortcomings of the CPS — (1) the underreporting of property income and other irregular sources of income; (2) the underreporting and misclassification of participation in major income security programs and other types of information that people generally find difficult to report accurately (for example, monthly detail on income earned during the year);

¹ This paper reports the general results of the research undertaken by Census Bureau staff. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau.

² Daniel Kasprzyk, SIPP Research and Coordination Staff, United States Bureau of the Census, Washington D.C. 20233

and (3) the lack of information necessary to analyze program participation and eligibility. Several features distinguish field tests of the ISDP from other data collections, particularly the CPS. They include: (1) interviews for the same persons were obtained at regular intervals within a year; (2) most types of income were reported on a monthly basis; (3) income was reported on an individual basis; (4) individuals were followed over the survey period to obtain data on changes in income and family composition; and (5) information was collected on special topics such as disability, child care, fertility, net worth, and taxes paid to provide insight into the context of program benefits, program dependency, and overall economic well-being. Because the ISDP was the predecessor to SIPP, many characteristics of the ISDP can be seen in the SIPP, including the survey design, content, and questionnaire format.

The SIPP began in October 1983 as an ongoing survey program with one sample panel of 21,000 households selected to represent the noninstitutional population of the United States. Each household is interviewed once every four months for approximately 2½ years; the reference period for the principal survey items is the 4 months preceding the interview. This interviewing plan results in eight interviews per household. Each year a new panel is introduced. This design allows cross-sectional estimates to be produced from the combined sample of 2 panels. Information concerning the SIPP design, content, and operations can be found in Nelson, McMillen and Kasprzyk (1985).

This paper reviews specific methodological, survey design, and statistical issues of concern to the program. The general categories of interest are: (1) questionnaire design; (2) data collection, including respondent rules, data collection mode, length of reference period, and rules for following movers; (3) concepts, design, and estimation; and (4) response error.

2. QUESTIONNAIRE DESIGN

The principal effort of the ISDP was directed to overcoming problems which resulted in underreporting and misclassification of income in the CPS March Supplement. In an ISDP field test, two questionnaire approaches were developed. For simplicity, one version may be referred to as the "short" form and the other as the "long" form. The short form approach attempted to gather income data directly while keeping respondent burden at a moderately low level. For each household member, questions were asked directly about the receipt of certain income types. If income were received, the amount received during the reference period was determined before proceeding to the next source of income.

The general strategy of the long form approach was to isolate events, experiences, and other attributes associated with the receipt of specific types of income. This form contained an extensive set of probes about the receipt of income and lengthy questions to ascertain income amounts. Amounts associated with specific income types were not obtained until all sources of income were determined.

The hypothesis tested was that the long form approach produces more complete and accurate reporting of income; Olson (1980) provides a summary of the analysis conducted on the two questionnaire formats. Several approaches to the analysis were implemented and are discussed in Olson's summary: (1) staff observation of training and interviewing; (2) debriefing sessions of interviewers and observers; (3) case-by-case reviews of completed questionnaires; (4) analysis of survey and item response rates; and (5) data analyses focussing on the quality of the data collected, and questionnaire edit failures, especially those associated with the inability of the interviewer to follow questionnaire skip patterns. The form adopted for further research and ultimately the SIPP was a variation of the long form. The long form was perceived

by both interviewers and respondents as less burdensome and also was shown to have higher income reporting rates.

An experiment with questionnaire formats was also included in the ISDP; this experiment contrasted a household screening format with a person-based format which had evolved from prior ISDP field tests. The household screening approach was based on a revised version of the questionnaire used in the April 1978 CPS Income Supplement Test and was intended to reduce burden by asking a single household respondent whether anyone in the household received a particular kind of income during the reference period. Each affirmative response was followed by a question to identify exactly which household member(s) received that type of income. Complete reciprocity for all household members was recorded before asking about amounts of income received by specific individuals. This approach was expected to reduce interview time without reducing data quality.

The approach above was contrasted with a person-based approach. Under this approach, questions on all sources of income were asked of the first household member, then repeated for the second, and so on. A separate form was filled out for each adult in a sample household, but extensive use was made of skip instructions and check items to reduce the number of questions asked of any one respondent.

Differences in the quality of the data obtained with the two questionnaire formats and differences in the interview times appeared slight. Large differences were not observed between the two approaches in estimates of income reciprocity rates, and in the incidence of "don't know" and "refusals." Interview time, expected to be significantly less under the household questionnaire approach, was about five minutes less per household and about three minutes less per person than the person approach. Since the household screening format did not offer a significant improvement over the person-based approach, this person-based format, with modest improvements and refinements, was adopted for SIPP.

Questionnaire design issues and discussions concerning data collection procedures continue to be part of the SIPP program. The general issue is whether interviews conducted without the use of responses from previous interviews (the so-called independent approach) produce better estimates than interviews conducted using the previous interview responses to remind respondents of earlier statuses (the so-called dependent interview approach). In the SIPP, a dependent approach is used to update income receipt patterns at each interview, but the approach has not been systematically evaluated.

A similar dependent approach to data collection is also possible with the data collected in the SIPP on personal net worth. These data are obtained at two points-in-time, one year apart. Specifically, data on asset and liability values, collected in Wave 4 of the 1984 Panel, were provided to one-half of the respondents interviewed in the Wave 7 interview. To examine differences between the dependent and independent approach, one half the sample in Wave 7 was provided information on asset and liability values collected in Wave 4, while the other half was not provided the previously reported information.

The rationale for this dependent or "feedback" approach was that respondents would provide more accurate estimates of change if they were first reminded of the amount they reported the previous year. If respondents know the amount of the change in asset values and were reminded of their beginning balance, then presumably their reporting of the current balance would be consistent with the true amount of change over the period. Lamas and McNeil (1987) analyze these data, but give no definite answer about the impact of the feedback approach since benchmark data are not available. They do, however, say that the dependent interview did not affect cross-sectional estimates and that the approach produced results consistent with expected differentials in net worth across subgroups. They also looked at micro-level changes in net worth

using only households with fully reported wealth data and found some evidence that the dependent interview reduced the estimates of the change in net worth.

The same questionnaire design issue, the dependent versus independent interview, has also occurred in the repeated measurement of industry and occupation. During the 1984 and 1985 SIPP panels these data were collected independently during each interview even though the individual had not changed employers. This procedure acknowledges the fact that an employee's duties may change from time to time and allows these changes to be recorded. Sufficient change in duties can result in a change in the person's occupation classification from interview to interview even though the employer has not changed.

The independent collection of industry and occupation data has, however, several problems. Undue variation in occupation classification can result when respondent descriptions of duties vary slightly or when the interpretation of the written description varies between the clerical staff members assigning the classification codes.

Research into this problem has provided some estimates of the number of times occupation and industry classifications change from interview to interview for persons with the *same* employer. Among individuals who reported the same employer during the first 12 months of the 1984 SIPP Panel, approximately 40 percent of these persons changed 3-digit occupation codes between two consecutive interviews and 20 percent changed 3-digit industry codes (Kalton, McMillen and Kasprzyk, 1986).

As a result, a modification was made in the 1986 SIPP Panel to reduce changes in occupation and industry codes resulting from random response error and clerical interpretation, and to reduce interview time. The modification introduces a "screener" question that asks if activities or duties have changed during the past 8 months. A negative response eliminates the detailed occupation and industry questions. The occupation and industry classifications are then derived from responses given in the previous interview.

It is important to note that while this change was made for the 1986 SIPP Panel, industry and occupation data from the 1985 SIPP Panel, collected during the same time period, were still collected independently each wave, giving rise to a natural experiment embedded in the two panels. These data have not yet been analyzed.

3. DATA COLLECTION

Four topics affecting data collection in the SIPP are discussed below: (1) respondent rules; (2) data collection mode; (3) length of reference period; and (4) rules for following movers.

Respondent Rules

When interviewing households with more than one member, a problem which must be addressed is the extent to which proxy responses are acceptable. Since not everyone may be present at the time of the interview, both time and money can be saved by asking another household member about persons who are not present. The difficulty with this is that along some dimensions of the survey instrument, the proxy report may result in less accurate data than the self-report. Kalton, Kasprzyk and McMillen (1988) provide a discussion of this issue in the context of panel surveys.

A formal test of respondent rules, conducted in the ISDP, compared the quality of reporting in a treatment group where proxy interviews are accepted from any household member who felt qualified to answer for a missing person with a treatment group where proxy interviews are not permitted except for extreme situations (respondent physically or mentally incapable,

unable to speak English, away from the household during the entire interviewing period, etc). About 85 percent of adults interviewed in the self-response rule households were self-respondents and about 65 percent were self-respondents in the usual or proxy response rule households. Thus, the implementation of the self-response rule resulted in approximately 20 percent more self-interviews than the other treatment (Coder 1980).

Refusal rates were slightly higher for the self-response treatment and the percent of households interviewed was slightly higher for the proxy response treatment. The differences, however, were too small to give insight into which rule should be preferred. Person noninterview rates in households where at least one other adult was interviewed were higher under self-response rules than under usual response rules. Differences between treatment groups in reported income reciprocity rates also appeared to be small and unaffected by the response rule, and combined "don't know" and "refusal" rates for income amounts of various income types were not consistently lower under the self-response mode.

Under the self-response rules, records were used more often by persons when answering wages and salary questions, and response rates for hourly wage rates were higher, but in general the evidence for either set of response rules was not conclusive. Thus, as a result of these findings, estimated costs for using a self-response rule (4-6 percent higher than the proxy rule), and the implementation of a "call back" procedure to obtain certain critical information unavailable at the time of the interview, the SIPP respondent rules now allow proxy interviews to be taken.

A related problem is the response rule for college students. Students are usually considered members of their parents' households until they establish a permanent residence elsewhere. Thus, the usual procedure for students living away from home while attending school is to treat them as household members who are temporarily absent and obtain proxy interviews from other members of their parents' household. In order to measure the accuracy of information taken from proxy interviews for students living away from home, one interview during an ISDP field test was first obtained by proxy at the parents' household and then by self-interview at the student's school residence. The results of this study are described by Roman and O'Brien (1984). The analysis presented is limited due to flaws in the administration and implementation of the test. The authors observed, however, that quite often a proxy cannot identify a particular source of student income and even if they can identify it, they are more likely to respond "don't know" to the particulars about that source. They also noted that the larger the income or expense, the better the proxy response becomes.

Data Collection Mode

The SIPP has conducted most interviews (approximately 95 percent) face to face (Kalton, McMillen, and Kasprzyk, 1986). Because of the rising costs of a face to face interviews, the Census Bureau is considering the possibility of conducting a substantially larger number of SIPP interviews by telephone.

As a result, a SIPP telephone interview pretest was conducted in June 1985 to assess the feasibility of "warm" telephone interviewing for SIPP — that is, telephone interviews for households which received a face to face interview at an earlier wave. The pretest was conducted in 2 of the Census Bureau's Regional Offices with a sample of 280 households. Refusal rates (about 2.5%) and noncontact rates (about 11%) were within staff's expectations. Item nonresponse rates showed no unexpectedly high nonresponse rates (U.S. Bureau of the Census 1986).

Following this, a SIPP National Telephone Test took place from August to November 1986 and February to April 1987; the purpose of the test was to study the large-scale use of warm telephoning in SIPP and to learn whether people are willing to furnish data by telephone for

2 interviews in a row. Households within 50 percent of the segments were designated as maximum telephone interview cases; the remaining 50 percent were maximum personal visit cases. Interviewers conducted almost all of the telephone interviews from their homes. Gbur and Durant (1987) report preliminary results from the first phase of the experiment. They indicate that household response rates did not seem to be seriously affected by the use of the telephone and person nonresponse rates were comparable by mode. Item nonresponse rates were only slightly affected by telephone interviewing. Additional results are forthcoming.

Length of Reference Period

The ISDP focussed on data collection techniques designed to improve the reporting of cash and noncash income, and as such the length of the reference period for most survey items was an important design decision.

This issue was addressed twice during the ISDP. First a single interview using a six month recall period was compared with two consecutive interviews, both using 3-month reference periods. Second, an experiment was conducted comparing reported property income amounts using a 3-month recall versus those with a 6-month recall period.

Olson (1980) describes some analyses conducted on the first experiment. Not surprisingly, using a 6 month recall period understates the proportion of income reported in earlier periods. This pattern held for a number of specific sources of income such as wages, Aid to Families with Dependent Children, and unemployment compensation. These findings though not definitive, support the presumption that longer recall periods increase chances of omission due to memory loss. Other analysis showed that the number of sources of income reported per household in the first three months of the six month reference period was lower than for the corresponding time using a three month reference period. Analyses of the second experiment were not conducted due to the withdrawal of funding for the development program.

The results of the first experiment along with the additional ISDP experience led to a four month recall period for the SIPP; this decision maintains cost at the appropriate budget level while trying to maintain satisfactory data quality.

Rules for Following Movers

An important design feature in the ISDP and now the SIPP is that all persons in a sample household at the time of the first interview remain in sample during the 2-½ year period of the panel; this rule holds even if one or more persons should move to a new address. For cost and operational reasons, face to face interviews are conducted at new addresses that satisfy some geographic constraint — in the ISDP, the address had to lie within 50 miles of an ISDP primary sampling unit, while in SIPP, the address must lie within 100 miles of a SIPP primary sampling unit.

For each panel a sample of addresses is selected and individuals are identified at these addresses at the time of the first interview. After the first interview, the sample is no longer address-based but rather person-based, consisting of all individuals enumerated during the first interview. These people and anyone with whom they share living quarters are interviewed in subsequent interviews.

During the ISDP two issues concerning movers were important: (1) the production of cross-sectional point in time estimates at each interview; and (2) the costs associated with following movers. Huang (1984) presents several unbiased base weights for cross-sectional estimates of the noninstitutionalized population when the sample contains movers. He associates observations at any given point in time with the known inclusion probabilities of the original sample

households. Two approaches are described: (1) a multiplicity approach, which depends on the number of ways that a new household can be included in the sample; and (2) a "fair share" approach which assumes all household members contribute equally to their household. The SIPP as well as the ISDP adopted the "fair share" approach.

The issue of costs was addressed by a "Mover's Cost Study". This study was to shed some light on the data collection costs resulting from following movers to their new addresses. White and Huang (1982) describe the study and provide some results based on the movers procedures adopted for the field test. They found that the number of eligible households for interview increased by 8.8 percent as a result of following movers during a one year time period; they also found that movers represented about 22 percent of the total sample after 15 months, and that during this period of time the number of interviewing hours increased by 7 percent and the number of miles charged by interviewers increased by 11.4 percent.

Jean and McArthur (1984) discuss data collection issues in the SIPP as they pertain to movers and offer recommendations to improve coverage in future SIPP panels. Kalton and Lepkowski (1985) also discuss the procedures for following movers in SIPP, and propose a research program aimed at measuring the extent of noncoverage from various sources and its concentration in particular subgroups. More recently, Jean and McArthur (1987), considering five waves of SIPP data, report that among persons who moved sometime after the first interview (that is, between Waves 2 and 5), 69 percent completed all 5 interviews, 23 percent did not complete the fifth interview, and 9 percent were interviewed in the fifth wave but were missing at least one intervening interview.

4. CONCEPTS, DESIGN AND ESTIMATION

During the ISDP and continuing with the SIPP program, significant research activity has taken place in the area of conceptualizing annual units of analysis using subannual data, and the statistical estimation of these concepts. The treatment of nonresponse in panel surveys has also been a topic of considerable research interest. Finally, estimation techniques to reduce sampling error and methods to sample subgroups have also been under study in the ISDP and SIPP programs.

Longitudinal Concepts

Annual family and household statistics are important indicators of the Nation's economic well-being. The SIPP collects subannual data, indeed monthly data, reflecting changes in the composition of households; these data allow the development of annual household statistics which reflect actual household composition experiences during the year, unlike current household statistics which simply ignore intrayear changes in household composition. The construction of annual units of analysis, whether they are households, families, or program units, raises methodological issues concerning longitudinal weights and imputation techniques. The main issue is, however, conceptual. Given intrayear composition change, when is it appropriate for annual measures to recognize change in household composition and when is it not? Put another way, how should households and families be defined which account for survey measurements at two or more points in time and which do not create serious conflicts with the traditional cross-sectional household and family constructs.

Analysts at the Census Bureau have given considerable thought to the question of defining households and families over time (McMillen and Herriot 1985; Citro 1985). Empirical research to examine several definitions of longitudinal households and measures of annual income status

and family type has been reported by Citro, Hernandez and Herriot (1986) and Citro, Hernandez and Moorman (1986). The empirical research emphasized four alternative concepts: (1) a household is the same over time if it has the same reference person; (2) a household is the same over time if it has the same principal person (this definition differs from the first in its treatment of married couple households for which the reference person may be either the husband or wife, but the principal person is always the wife); (3) a household is the same over time if it has the same reference person and is the same family type over time; and (4) a household continues over time if it has the same reference person, is the same family type, and has the same membership size.

This research has provided preliminary indications that the choice of definition does not appreciably affect annual measures of low income status or of households by type. If this finding does not change after additional research, considerations, such as ease of implementation and operational simplicity, will be the determining factors in the use of a longitudinal household definition.

Statistical Estimation for Longitudinal Concepts

Research on estimation for longitudinal concepts has proceeded along two paths — longitudinal person estimation and longitudinal household (family or program unit) estimation. The work on person estimation includes the calculation of selection probabilities to yield unbiased longitudinal estimates of individual characteristics and the use of controls in additional stages of estimation (Judkins *et al.*, 1984). A refinement of this work and a description of the method proposed to produce longitudinal weights for person analysis covering the first three SIPP interviews has been reported by Kobilarcik and Singh (1986).

Kobilarcik and Singh define the longitudinal universe as the noninstitutional population (excluding military barracks) on December 1, 1983, the midpoint of the Wave 1 interview months. The sample from the longitudinal universe consists of eligible persons living in the selected living quarters at the time of the first interview. "Interviewed" persons for purposes of this estimation procedure are those who responded to each of the first three SIPP interviews, and who during the first interview lived in a household in which all eligible members responded to the interview, and those who resided in a Wave 1 interviewed household, but during the second or third interview died or moved outside the geographic boundaries of the survey.

Thus, noninterviewed persons in the estimation procedure are those who at the time of the first interview lived in a household in which at least one household member failed to respond to the first interview, and those who resided in a Wave 1 interviewed household but failed to respond at the second and/or third interview. All persons classified as interviewed are assigned positive weights. Weights for this universe are derived in the usual way, using the reciprocal of the probability of selection, calculating an adjustment for noninterviews, and adjusting to demographic population controls. The nonresponse adjustment has two phases, an adjustment first for household nonresponse and then for person nonresponse, the latter using information collected during the first interview.

The topic of longitudinal household (family or program unit) estimation is also under study. Several approaches to this issue were reported by Ernst, Hubble and Judkins (1984) and more recently by Ernst (1988). The latter work describes why weighting by the reciprocal of the probability of selection does not, in general, work for longitudinal household and family estimates, and presents a class of weighting procedures which can accomplish this task. He, furthermore, describes the difficulties that can arise in applying these weighting procedures because the information necessary to create the weight may not be available. Ernst also presents conditions which,

if satisfied, by the longitudinal concept, are sufficient for there to exist a weighting procedure that avoids these problems. Finally, he discusses procedures for adjusting longitudinal concepts for nonresponse and for controlling demographic variables to independent estimates.

Nonresponse and Imputation

For longitudinal surveys such as those of the ISDP and the SIPP, the problems of refusal and selective nonresponse are compounded by cumulative losses in responses over the course of the panel. Therefore, an important aspect of both the ISDP and SIPP work has been the study of methods for compensating for nonresponse. To that end, Kalton (1983) reviewed procedures used in survey research. Imputation procedures were also discussed by Kalton and Kasprzyk (1982, 1986), where bias and variance properties for several classes of procedures are summarized.

SIPP data can be treated as both cross-sectional and longitudinal. Procedures to compensate for unit nonresponse in the SIPP as well as other Census Bureau surveys are described in Chapman, Bailey and Kasprzyk (1986). Complications arising in the treatment of unit nonresponse in a multi-interview survey are described. In a panel survey, however, nonresponse may also occur, as item nonresponse, where a unit takes part in the survey but does not provide answers to all items, and as wave nonresponse where a unit provides data for some, but not all of the interviews.

Heeringa and Lepkowski (1986) describe general classes of longitudinal imputation methods which might be considered as an alternative to a cross-sectional hot deck imputation approach. They also empirically compare a simple longitudinal imputation method, longitudinal direct substitution, where the value of a nonmissing item is substituted from one time period to another when the same item is missing, with a cross-sectional hot deck scheme. Not surprisingly, they demonstrate that the direct substitution method for longitudinal imputation understates change. They concluded, however, that this may be preferable to the gross overstatement of change resulting from the use of the cross-sectional hot deck method.

Panel surveys have an additional type of missing data problem called wave nonresponse. The amount of missing data for an individual with wave nonresponse is typically greater than that encountered for records with item nonresponse. Data available from completed waves of interviewing provide more detailed information about the nonresponding unit than is available for total nonrespondents. Thus, nonresponse compensation strategies may include weighting, imputation, or a combination of both. Kalton, Lepkowski and Lin (1985) discuss this issue and empirical findings in the context of the ISDP. This work made it clear that the choice between weighting and imputation for missing data of this type is far from obvious. Kalton (1986) and Kalton and Miller (1986) further refine the understanding of this problem and conclude that imputation can distort some forms of estimates and that weighting may be the preferred solution for large subclasses when the reduction in effective sample size is tolerable. They caution, however, that imputation may be better for estimates based on small subclasses when the loss of sample is important. In the case of a three interview longitudinal SIPP file the difference in sample size between weighting and imputation is not substantial, and consequently the weighting approach is the safer general purpose solution. Finally, Lepkowski (1988) after further empirical research concludes that a specific strategy for wave nonresponse can only be developed after consideration of such factors as the major survey design objectives, the panel design, and the distribution of wave nonresponse patterns. He provides criteria to be considered in developing missing data strategies and concludes that weighting strategies appear to be preferable for compensating for wave nonresponse.

Sampling Error Reduction through Estimation Techniques

Two methods for reducing sampling error through estimation techniques are under study: composite estimation and the use of administrative records in SIPP estimation.

Composite estimation is a technique that combines estimates from the current and previous time periods with the goal of improving the precision of survey estimates by taking advantage of the correlations between responses for the same analytic units at different time periods. Composite estimation is particularly effective when the correlations are high, which is likely to be the case for many important data items in the SIPP. Chakrabarty (1986) has conducted a preliminary review of the types of composite estimates appropriate for the SIPP data structure. The content of the survey has not been sufficiently stable during the first few years of the SIPP to seriously consider adoption of a composite estimator.

Another approach to variance reduction is through the use of administrative records for post-stratification. Currently, cross-section estimation procedures for SIPP make use of a second-stage adjustment to increase the precision of estimates by ratio adjusting collection month and reference month estimates to population estimates. However, the Census Bureau has access to some Internal Revenue Service and Social Security Administration files which can be used to produce detailed age, race, and sex distributions by adjusted gross income. The issue, which we have just begun to explore, is how these administrative data can be used for post-stratification to improve estimates of mean and median personal and household income as well as the estimates of the deciles of the personal and household income distribution. The basic question under study is the magnitude of the reduction in variances of these estimates achieved through such a procedure. Fay and Huggins (1988) will provide some indications.

Sampling for Special Subpopulations

Subgroups of the population are often cited as being more affected by governmental policy than others — the population of persons in poverty, the aged, the Blacks, Hispanics, and participants of Federal income security programs. Early design goals of the ISDP emphasized a concern for improving the reliability of subpopulation estimates. This was exhibited in the emphasis placed in the ISDP on sampling from administrative program lists. Thus, samples were oftentimes drawn from lists of current participants of Federal or state administered programs (Kasprzyk 1983; Bowie and Kasprzyk 1987).

A Census Bureau Working Group analyzed subsampling (screening) proposals for oversampling special populations. The issue studied concerned the reliability of estimates when different subsampling schemes are introduced. Subsampling characteristics based on income and demographic variables were identified and estimates of reliability for different subsampling rates and characteristics were calculated (U.S. Bureau of the Census 1985).

This group concluded that subsampling proposals, for a general purpose income survey like the SIPP, provided only modest gains in precision for low-income items and did not outweigh the disadvantages, which included an increase in the complexity of the operation, the loss of a self-weighting design, and large decreases in precision for the middle income items.

5. RESPONSE ERROR

Response error is one aspect of a more general problem, nonsampling error, discussed by Kalton, Kasprzyk and McMillen (1988). Response error occurs when incorrect data are recorded on the questionnaire. This can occur for a variety of reasons, such as a faulty questionnaire, memory errors, inappropriate respondents, etc. In this section we briefly describe a response error issue with the SIPP gross flow data and a record check study aimed at providing insight into a better understanding of response errors in general.

SIPP Gross Flow Data

Analysis of program data on a month-to-month basis in ISDP revealed a tendency for reported program turnover to occur between waves of interviewing more often than within the wave (Moore and Kasprzyk 1984). Analysis using the SIPP data (Burkhead and Coder 1985) covering month-to-month changes in receipt of income benefit amounts for a 12 month period focussed on changes occurring between the last month of one reference period and the first months of the succeeding reference period. The results using SIPP and ISDP data are similar, where an uneven pattern of change is observed and this pattern is clearly associated with the interviewing scheme. Gross changes are significantly higher between the last month of one reference period and the first month of the next. Hill (1987) used monthly data from the 1984 and 1985 waves of the Panel Study of Income Dynamics (PSID) to investigate the extent and determinants of excessive change between waves relative to measured change within waves of a panel survey. He found that in spite of different question sequences, and recall periods, between wave transitions dominate the within wave transitions in the PSID just as they do in the SIPP. The main causes for the problem are not known, but questionnaire wording/design, respondent recall error, and the interaction between these two factors seem likely.

Weidman (1986) did an empirical analysis to look for obvious relationships between respondent characteristics and changes in receipt status of a number of income types. He did not detect any relationship between gross change distributions, self/proxy status and nine demographic variables (age, race, sex, education, marital status, household size, tenure, relationship to reference person, and size of metropolitan area) for consecutive months, but did note that more transitions occur when some of the data are imputed. The absence of any notable relationships indicates a need for exploring other ways to understand this problem.

Interest in gross flow estimates remains high. Hubble and Judkins (1986) developed a model to estimate biases in gross flows estimates resulting from response errors, the parameters of which are estimated using SIPP response error rates and the ratios of within-wave and between-wave gross flow estimates. Several strong assumptions, as well as a reinterview program which produces accurate reinterview data on gross flows within the period, are necessary. Weidman (1987) presents linear models that try to represent the relationships between observed and actual transitions. The models are admittedly oversimplified using only survey reported data, but nevertheless, illustrate the need to obtain more information about the SIPP error structure in reporting receipt of benefits from government transfer programs.

SIPP Record Check Study

One way to study the SIPP error structure in reporting receipt of program benefits and amounts is to develop validation studies of items common to both survey records and administrative records. The SIPP program has initiated such a study to investigate response quality issues.

The goal is the improved understanding of the quality of the SIPP data and, ultimately, the development of quantitative estimates of response and nonresponse errors in order to adjust the survey data or modify survey procedures to obtain better quality data. The research questions addressed in this study include: (1) the quality of the respondent reports of receipt of program benefits for a variety of state and Federally administered transfer programs; (2) the quality of benefit dollar amount reporting for these programs; (3) demographic correlates of report quality; (4) extent of misclassification errors; (5) the effects of self-proxy respondent status on report quality; and (6) between wave reciprocity turnover effects. Four state administered programs and six Federally administered programs are included in the study. Moore and Marquis (1987) provide very preliminary results, suggesting that reporting problems

are different for the Aid to Families with Dependent Children (AFDC) and the Food Stamp Programs, the former having a net under-reporting as well as a time placement problem for reporting a transition in program status while the latter having only a time placement problem.

6. CONCLUSION

As in all large-scale continuing survey programs, research is needed to improve understanding of the effects of survey methods on the data collected. A survey, like the SIPP, which is complex in its implementation requires a commitment to understanding the measurement process. The wide range of topics discussed above — collection, longitudinal concepts and estimation, and response error — illustrate where the interest and emphasis was placed during the development program and the first few years of the SIPP program.

ACKNOWLEDGEMENTS

The author gratefully acknowledges the helpful comments of the referees and the secretarial and clerical help of Ms. Hazel Beaton in the preparation of this manuscript.

REFERENCES

- BOWIE, C., and KASPRZYK, D. (1987). A review of the use of administrative records in the Survey of Income and Program Participation. *SIPP Working Paper Series No. 8721*, Washington, D.C.: U.S. Bureau of the Census.
- BURKHEAD, D., and CODER, J. (1985). Gross changes in income recipiency from the Survey of Income and Program Participation. *Proceedings of the Social Statistics Section, American Statistical Association*, Washington, D.C., 351-356.
- CHAKRABARTY, R. (1986). Composite estimation for SIPP: A preliminary report. *SIPP Working Paper Series No. 8610*. Washington, D.C.: U.S. Bureau of the Census.
- CHAPMAN, D., BAILEY, L., and KASPRZYK, D. (1986). Nonresponse adjustment procedures at the U.S. Bureau of the Census; *Survey Methodology*, 12, 161-180.
- CITRO, C. (1985). Alternative definitions of longitudinal households in the Income Survey Development Program: Implications for annual statistics. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 381-386.
- CITRO, C., HERNANDEZ, D., and HERRIOT, R. (1986). Longitudinal household concepts in SIPP: Preliminary results. *Proceedings of the U.S. Bureau of the Census Second Annual Research Conference*, 598-619.
- CITRO, C., HERNANDEZ, D., and MOORMAN, J. (1986). Longitudinal household concepts in SIPP. *Proceedings of the Social Statistics Section, American Statistical Association*, Washington, D.C., 361-366.
- CODER, J. (1980). Some results from the 1979 Income Survey Development Program research panel. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 540-545.
- ERNST, L. (1988). Weighting issues for longitudinal household and family estimates. In *Panel Surveys* (Eds. D. Kasprzyk, G. Duncan, G. Kalton, and M.P. Singh), New York: John Wiley, forthcoming.
- ERNST, L., HUBBLE, D., and JUDKINS, D. (1984). Longitudinal family and household estimation in SIPP. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 682-687.

- FAY, R.E., and HUGGINS, V.J. (1988). Use of administrative data in SIPP longitudinal estimation. Paper to be presented to the Section on Survey Research Methods, American Statistical Association, August 1988.
- GBUR, P., and DURANT, S. (1987). Testing telephone interviewing in the Survey of Income and Program Participation and some early results. Paper presented at the International Symposium on Telephone Survey Methodology, 1987, Charlotte, North Carolina.
- HEERINGA, S., and LEPKOWSKI, J. (1986). Longitudinal imputation for the SIPP. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 206-210.
- HILL, D. (1987). Response errors around the seam: Analysis of change in a panel with overlapping reference records. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D. C., 210-215.
- HUANG, H. (1984). Obtaining a cross-sectional estimate from a longitudinal survey: Experience of the ISDP. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 670-675.
- HUBBLE, D., and JUDKINS, D. (1986). Measuring the bias in gross flow estimates in the presence of auto-correlated response errors. *Proceedings of the Section on Survey Research Methods, American American Statistical Association*, Washington, D.C., 237-242.
- JEAN, A., and McARTHUR, E. (1987.) Tracking Persons Over Time. *SIPP Working Paper Series No. 8701*, Washington, D.C.: U.S. Bureau of the Census.
- JEAN, A., and McARTHUR, E. (1984). Some data collection issues for panel surveys with application to the Survey of Income and Program Participation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 745-750.
- JUDKINS, D., HUBBLE, D., DORSCH, J.R., McMILLEN, D., and ERNST, L. (1984). Weighting of persons for SIPP longitudinal tabulations. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 676-681.
- KALTON, G. (1986). Handling wave nonresponse in panel surveys. *Journal of Official Statistics*, 2, 303-314.
- KALTON, G. (1983). *Compensating for missing survey data*. Survey Research Center, University of Michigan.
- KALTON, G., and KASPRZYK, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1-16.
- KALTON, G., and KASPRZYK, D. (1982). Imputing for missing survey responses. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 22-31.
- KALTON, G., KASPRZYK, D., and McMILLEN, D.B. (1988). Nonsampling errors in panel surveys. In *Panel Surveys* (Eds. D. Kasprzyk, G. Duncan, G. Kalton, and M.P. Singh), New York: John Wiley, forthcoming.
- KALTON, G., and LEPKOWSKI, J. (1985). Following rules in SIPP. *Journal of Economic and Social Measurement*, 13, 319-329.
- KALTON, G., LEPKOWSKI, J., and LIN, T. (1985). Compensating for wave nonresponse in the 1979 ISDP research panel. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 372-377.
- KALTON, G., McMILLEN, D., and KASPRZYK, D. (1986). A review of nonsampling error issues in the Survey of Income and Program Participation. *Proceedings of the U.S. Bureau of the Census Second Annual Research Conference*, 147-165.
- KALTON, G., and MILLER, M. (1986). Effects of adjustments for wave nonresponse on panel survey estimates. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 194-199.

- KASPRZYK, D. (1983). Social Security number reporting, the use of administrative records, and the multiple frame design in the Income Survey Development Program. In *Technical, Conceptual, and Administrative Lessons of the Income Survey Development Program* (ISDP), (Ed. M. David), Washington, D.C.: Social Science Research Council, 171-198.
- KOBILARCIK, E., and SINGH, R. (1986). SIPP longitudinal estimation for persons' characteristics. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 214-219.
- LAMAS, E.J., and McNEIL, J.M. (1987). An analysis of the SIPP asset and liability feedback experiment. *Proceedings of the Social Statistics Section, American Statistical Association*, 194-199.
- LEPKOWSKI, J. (1988). The treatment of wave nonresponse in panel surveys. In *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton, and M.P. Singh). New York: John Wiley, forthcoming.
- McMILLEN, D.B., and HERRIOT, R. (1985). Toward a longitudinal definition of households. *Journal of Economic and Social Measurement*, 13, 349-360.
- MOORE, J.C., and KASPRZYK, D. (1984). Month-to-month reciprocity turnover in the ISDP. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 726-731.
- MOORE, J.C., and MARQUIS, K. (1987). Using administrative data to evaluate the quality of survey estimates. Paper presented at the International Symposium on the Statistical Uses of Administrative Data, 1987, Ottawa, Canada.
- NELSON, D., McMILLEN, D., and KASPRZYK, D. (1985). An overview of the Survey of Income and Program Participation: Update 1. *SIPP Working Paper Series No. 8401*. Washington, D.C.: U.S. Bureau of the Census.
- OLSON, J. (1980). *Reports from the Site Research Test*, (Ed. J. Olson). Office of the Assistant Secretary for Planning and Evaluation, Department of Health and Human Services, United States.
- ROMAN, A.M., and O'BRIEN, D.V. (1984). The student follow-up investigation of the 1979 ISDP. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 732-737.
- U.S. Bureau of the Census (1986). Additional results from the SIPP telephone test. Memorandum from J. Coder to A. Norton, April 9, 1986.
- U.S. Bureau of the Census (1985). SIPP research on SIPP oversampling/subsampling. Memorandum from R. Singh to G. Shapiro and D. Kasprzyk, August 12, 1985.
- WEIDMAN, L. (1987). Examination of relationships between actual and reported changes in the SIPP. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 216-220.
- WEIDMAN, L. (1986). Investigation of gross changes in income reciprocity from the Survey of Income and Program Participation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 231-236.
- WHITE, G., and HUANG, H. (1982). Mover follow-up costs for the Income Survey Development Program. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 376-381.
- YCAS, M., and LININGER, C. (1981). The Income Survey Development Program: Design features and initial findings. *Social Security Bulletin*. Baltimore, Md.: Social Security Administration, 44.

Personal Computer Variance Software for Complex Surveys

DAN SCHNELL, WILLIAM J. KENNEDY, GARY SULLIVAN,
HEON JIN PARK and WAYNE A. FULLER¹

ABSTRACT

A personal computer program for variance estimation with large scale surveys is described. The program, called PC CARP, will compute estimates and estimated variances for totals, ratios, means, quantiles, and regression coefficients.

KEY WORDS: Survey sampling; Variance estimation; Survey software.

1. INTRODUCTION

The analysis of survey data typically involves a large number of observations and relatively complex variance calculations. Recent developments in personal computers have made possible the use of such computers to process data from complex surveys. We describe a personal computer program for survey data analysis prepared at Iowa State University.

The project to develop statistical software for variance estimation on the personal computer was a joint undertaking between Iowa State University and the International Statistical Programs Center of the U.S. Census Bureau. The objective of the Census Bureau was to provide developing countries with software that can be used locally to process survey data collected locally. The Iowa State University project on variance estimation was part of a larger Census Bureau undertaking that included the development of software for survey management, data editing and tabulation.

Beginning in the early 1970's, based on the work of Hidiroglou (1974) and Fuller (1975), a program was developed at Iowa State University for the computation of regression coefficients and the estimated covariance matrix of the coefficients for survey data. The program, called SUPER CARP, was later expanded to include total estimation, ratio estimation, subpopulation statistics, two-way tables and two stage samples. The last revision of SUPER CARP took place in 1980. SUPER CARP furnished the starting point for software development on the personal computer. Because of its ancestry, the personal computer program was called PC CARP.

2. PROGRAM CAPABILITY

PC CARP was designed for the IBM PC, IBM PC/XT, IBM PC/AT and compatible machines. At least 410K bytes of memory and a math coprocessor are required.

PC CARP is capable of handling both large and small data sets with equal ease and efficiency. The program sets no limit on the number of strata or clusters that can appear in a data set and can accept up to 50 input variables at a time. The program accepts disk data files in either fixed or free format.

The program can be used to compute variances for one or two stage samples with finite population correction terms included. For samples with more than two stages, finite population

¹ Dan Schnell, Centers for Disease Control, 1600 Clifton Road, NE, Atlanta, Georgia 30333. William J. Kennedy, Gary Sullivan, Heon Jin Park and Wayne A. Fuller, Department of Statistics, Iowa State University, Ames, Iowa 50011, United States.

corrections are only available at two levels. For two-stage samples, the program computes within cluster sampling rates from the stratum sampling rates and the individual observation weights.

Typically, each observation in the data file will contain stratum identification, cluster (primary sampling unit) identification, and a weight where the weight is the inverse of the selection probability. The user may or may not elect to enter first stage sampling rates. For simple designs, such as simple random sampling, not all of this information is required. In such cases reduced data input is possible.

If stratification is present, the program requires that all observations belonging to the same stratum be grouped together. If clustering is present, all observations belonging to the same cluster must be grouped together.

Table 1 contains a description of the types of statistics available to the user of PC CARP. In addition to the items of Table 1, supplements are available for estimation of the logistic function and for post stratified samples. These supplements are discussed in Section 4.4 and Section 4.5. An "X" in the column headed "Cov. matrix" means that the covariance matrix of a vector of estimates of the type listed on the left can be obtained. The standard error is computed for all statistics, but the covariance matrix of a vector is available for only a restricted set. Also, the coefficient of variation is computed for many statistics. The design effect, denoted by DEFF, is available as an option for many of the statistics. See Kish (1965) for a description of the design effect.

Table 1
Analysis Capabilities of PC CARP

Analysis	Coeff. var.	Cov. marix	Design effect	Comments
<i>Population Analyses</i>				
Total Estimation	X	X	X	50 variables maximum
Ratio Estimation	X	X	X	50 variables maximum without covariances 15 with covariances
Difference of Ratios			X	15 variables maximum
<i>Stratum Analyses</i>				
Totals	X		X	50 variables maximum
Means	X		X	50 variables maximum
Proportions	X		X	50 variables maximum
<i>Subpopulation Analyses</i>				
Totals	X		X	Crossed classif.
Means	X		X	Multiple variables
Proportions	X		X	Crossed classif. Multiple variables
Ratios	X		X	Crossed classif. Multiple variables
<i>Other Analyses</i>				
Two-way Table		X		50 cells maximum proportionality test
Regression		X		50 variables maximum Multiple d.f. tests Y-hat, residuals
Univariate			X	Multiple variables, empirical CDF, quantiles

The population (Total, Ratio and Difference of Ratios) analyses and stratum analyses are performed in a straightforward manner. Some details pertaining to Subpopulation Analyses, the Two-Way Table, Regression Analysis, and Univariate Analysis are presented in Section 4.

The subpopulation analyses give the user the option of crossing classification variables. This allows the user to create new classification structures from two or more input classification variables. For example, suppose the input data includes the classification variables age, sex and education with six, two and five levels, respectively. Then, by crossing age with sex with education, a new classification structure with 60 levels is produced. The user may obtain estimates for any number of dependent variables under this classification structure.

The Two-way Table analysis is defined by two classification variables and a dependent variable. More than one dependent variable can be specified for a pair of classification variables. Tables of cell totals, of proportions based on row totals, of proportions based on column totals, and of proportions based on the grand total are computed for each dependent variable. Standard errors are computed for all estimators and a test statistic for the hypothesis of proportionality is output. The test statistic is based on a Satterthwaite approximation to the distribution of the Pearson chi-square statistic. Also see Rao and Scott (1984).

The weighted least squares regression analysis computes coefficient estimates, and an estimated variance-covariance matrix which takes into account the sample design. These calculations are given in Fuller (1975) and outlined in Hidioglou *et al.* (1980). Multiple degrees of freedom F-tests for sets of coefficients and the usual t-statistics are available. The user also has the option of obtaining residuals and predicted values.

The Univariate analysis provides statistics that describe the distribution of a variable. The user specifies the variable of interest and identifies a subpopulation by specifying a category of a classification variable. Thus, the user might elect to obtain statistics for the personal income of individuals in the professional category of the occupation classification. Estimates of the subpopulation mean, variance, distribution function, quantiles and interquartile range are produced.

3. PROGRAM DETAILS

PC CARP is written almost entirely in FORTRAN, the most widely known scientific programming language, and the IBM Professional FORTRAN compiler was selected for the project. A small portion of the code — some sections of the user interface — is written in IBM Assembly language.

Two concerns at the program development stage were to provide a friendly user interface and to minimize the number of passes through the data. The interface was made user friendly by implementing an interactive, screen oriented response system. A single pass algorithm for variance estimation of simple statistics minimized the amount of reading from data files. Most estimators and their variances are obtained in a single pass through the data.

Estimators can be computed for the total population, for each stratum, or for specified subpopulations. For the most part, the estimators are functions of weighted sample totals. For example, to compute the estimators of the ratios, $R_1 = Y_1/X_1$ and $R_2 = Y_2/X_2$, one accumulates totals for Y_1 , X_1 , Y_2 , and X_2 . If the estimate is for the entire population, these totals are accumulated in one pass through the data. Totals for stratum estimates can be accumulated, combined if necessary, and output stratum by stratum. Since the data are grouped by strata, stratum totals can also be obtained in one pass for any number of strata. Subpopulation estimators may require more than one pass through the data if the number of categories defined by the classification structure is large. The Regression and Univariate analyses require two passes through the data.

The estimators, with the exception of totals, are nonlinear functions of weighted sample moments. It follows that a method appropriate for a nonlinear function must be used to estimate the variance of the approximate distribution of such estimators. See Wolter (1985) for a discussion of variance estimation for complex surveys. The Taylor method (method of statistical differentials) is the method of variance estimation used in PC CARP. Generally, the Taylor method has been shown to be equal to or superior to other variance estimation methods for the statistics, such as ratios, under consideration. See, for example, Frankel (1971). The Taylor variance of the ratio estimator is given in such standard texts as that of Cochran (1977) and the Taylor variance of a regression coefficient is given by Fuller (1975).

The value of the estimator and its estimated variance can, in most cases, be computed in the same pass. This is because the first order Taylor approximation to the variance can be expressed in terms of the variances of totals. For example, the first order Taylor approximation to $\hat{R} = \hat{Y}/\hat{X}$ is

$$\hat{R} \doteq R + X^{-1}(\hat{Y} - R\hat{X}),$$

where $R = Y/X$ is the ratio of the true totals. It follows that the estimated variance of a ratio $\hat{R} = \hat{Y}/\hat{X}$ can be computed from the estimated variance of the totals of Y , X , and $(Y - X)$. Similarly, the estimated covariance matrix for $\hat{R}_1 = \hat{Y}_1/\hat{X}_1$ and $\hat{R}_2 = \hat{Y}_2/\hat{X}_2$, can be computed from the estimated variances of the totals of the ten quantities Y_1 , X_1 , $(Y_1 - X_1)$, Y_2 , X_2 , $(Y_2 - X_2)$, $(Y_1 - Y_2)$, $(Y_1 - X_2)$, $(Y_2 - X_1)$, and $(X_1 - X_2)$.

The algorithm used for the calculation of the weighted mean and weighted sums of squares and cross products matrices is described in Herraman (1968). For sample values X_i and corresponding weights $\{W_i\}$, the sequence of weighted means, \bar{X}_K , and weighted corrected sum of squares, S_K , is computed as

$$\bar{X}_K = \bar{X}_{K-1} + a_K d_K \quad \text{and} \quad S_K = S_{K-1} + D_K - D_K a_K,$$

where $d_K = X_K - \bar{X}_{K-1}$, $a_K = W_K(\sum_{i=1}^K W_i)^{-1}$, and $D_K = d_K^2 W_K$.

Up to three different variance quantities can be accumulated concurrently for any given estimator. These are the first stage variance component, the optional second stage variance component and the optional simple random sampling variance used in the computation of the design effect. Computing all variance quantities in a single pass through the data requires a large amount of array space. However, when working with large samples, the elimination of entire passes through the data out-weighs the use of additional memory.

The program routinely performs checks to avoid computational errors such as division by zero. For example, if the user enters a data set with only one cluster in a stratum, the program will assign zero variance to the stratum, complete the calculations, and print an error message identifying the stratum with a single cluster.

The error handling system was constructed to avoid program termination caused by user misspecifications that could be easily corrected. Checks for omitted responses, improper file names and invalid analysis variable specifications are included in the program. If such an error is detected, PC CARP permits the user to re-enter information or to exit the program.

Program accuracy was assessed by constructing examples and comparing results with those obtained using the mainframe program SUPER CARP. The data set of Longley (1967) was used to evaluate the accuracy of the regression program. Additional checks were made using PROC MATRIX of the SAS package. See Barr, et al. (1979). PC CARP numerical accuracy was found to be at the same level as the mainframe packages. Internal consistency of PC CARP was also verified by computing equivalent estimators using different options, e.g., by computing a subpopulation mean using the subpopulation option and using the ratio option.

When information is needed by PC CARP, the user receives a full screen of short response questions along with detailed instructions. The first set of screens displayed to the user ask for information pertaining to data organization and location. "Help" and "Go Back" options are available at many places.

The second phase of program execution is Analysis Specification. In this phase the user chooses the type of analysis, options for that type of analysis, and the analysis variables. Any number of analyses can be performed using the data specified in phase one.

4. SPECIAL FEATURES

4.1 Two Way Table

As described in Section 2, this option automatically provides the user with four tables, where the entries are determined by the type of marginal control exercised in constructing the table.

We outline the procedure used to construct the table of cell proportions and the estimated covariance matrix of the proportions. Suppose the table has R rows and C columns and let \hat{Y}_{rc} be the estimated total for the rc -th cell. Let \hat{Y} be the RC -dimensional column vector of cell totals, created by listing the columns of totals one beneath the other beginning with the first column. Let

$$\hat{Y}_{..} = \sum_{r=1}^R \sum_{c=1}^C \hat{Y}_{rc},$$

$$\hat{P}_{rc} = \hat{Y}_{..}^{-1} \hat{Y}_{rc}$$

be the estimated population total and the estimated cell proportion for cell rc , respectively. Let \hat{P} be the RC -dimensional column vector, analogous to \hat{Y} , composed of the RC values \hat{P}_{rc} , arranged by column. The estimated covariance matrix for \hat{P} is

$$\hat{V}_{PP} = \hat{Y}_{..}^{-2} [I_{RC} - (\hat{P} \otimes J'_{RC})] \hat{V}_{YY} [I_{RC} - (\hat{P} \otimes J'_{RC})]',$$

where \hat{V}_{YY} is the estimated covariance matrix of the vector of cell totals \hat{Y} , I_{RC} is the identity matrix of dimension RC , and J_{RC} is an RC -dimensional column vector of ones.

The matrix \hat{V}_{PP} is used to compute the test statistic for the hypothesis of proportionality. The null hypothesis for the test is the hypothesis that the interior entries in the population table are the products of the marginal proportions. See Rao and Scott (1984) for a discussion of tests for such hypotheses. The test in PC CARP is based on a Satterthwaite approximation to the distribution of the Pearson chi-square statistic constructed as if the proportions were multinomial proportions. The approximation is valid for any analysis variable.

4.2 Quantile Estimation

Among the statistics produced by the univariate option are estimates of quantiles and an estimator of the standard error of the quantiles. The first step in the computation of quantiles is the construction of an estimator of the cumulative distribution function. In a first pass through the data the range of observations, the sample mean, and the sample standard deviation are constructed. Also, the three largest observations and the three smallest observations are identified.

The estimated cumulative distribution function is defined by

$$\hat{F}_Y(x) = \left(\sum_{t=1}^m w_t Z_{St} \right)^{-1} \sum_{t=1}^m w_t Z_{St} I_Y(x),$$

where the summation is over the m elements in the sample, w_t is the sample weight, Z_{St} is an indicator function that is one if the observation is in the subpopulation of interest and zero otherwise, and $I_Y(x)$ is one if $Y < x$ and is zero otherwise. The range of the variable is divided into 100 intervals and the cumulative distribution function is estimated at the 101 values defined by this subdivision.

The covariance matrix for the estimated distribution function evaluated at 25 points, $j = 1, 5, \dots, 96$, is estimated. The estimated standard errors are smoothed with a three point moving average and interpolation is used to obtain an estimated standard error for each of 101 points of the estimated distribution function. Linear interpolation is used to create an estimated distribution function that is monotone increasing. Using the smoothed standard errors, a monotone increasing upper bound and monotone increasing lower bound that form a pointwise 95% confidence interval for the distribution function are established. These bounds are then inverted to form 95% confidence interval for the quantiles. The interquartile range and its standard error are also estimated.

The quantile estimation is based on a theory that assumes the existence of an underlying superpopulation distribution function with a positive density. See Francisco (1987) for theoretical details and Park (1987) for computational aspects.

4.3 Regression Estimation

Estimates of the coefficients of a linear regression model are computed by the method of weighted least squares. Using the procedure given in Fuller (1975), an estimator of the covariance matrix of the coefficient vector is computed, taking into account the sample design.

The coefficient vector is estimated by

$$\hat{b} = (X'WX)^{-1}X'WY,$$

where X is the $n \times p$ matrix of independent variable values, \hat{Y} is the n -dimensional vector of dependent variable values, W is a matrix with the observation weights on the diagonal and zeros elsewhere, and n is the total number of observations. The variance of \hat{b} is estimated by

$$\hat{V}(\hat{b}) = (X'WX)^{-1}\hat{G}_W(X'WX)^{-1}.$$

The matrix \hat{G}_W is

$$\hat{G}_W = C \sum_{i=1}^L h_i \sum_{j=1}^{n_i} (\hat{d}_{ij} - \bar{d}_{\cdot i})(\hat{d}_{ij} - \bar{d}_{\cdot i})',$$

where

$$\hat{d}_{ij} = \sum_{k=1}^{m_{ij}} X_{ijk} \hat{v}_{ijk} W_{ijk},$$

$$\hat{v}_{ijk} = Y_{ijk} - \hat{b}' X_{ijk},$$

$$\bar{d}_{i.} = n_i^{-1} \sum_{j=1}^{n_i} \hat{d}_{ij},$$

$h_i = (n_i - 1)^{-1} n_i$, $C = (n - p)^{-1} (n - 1)$, m_{ij} is the number of elements in cluster j of stratum i , n_i is the number of clusters in stratum i , n is the total number of observations, L is the number of strata, and p is the number of coefficients estimated. The variance estimator differs from the usual weighted least squares variance estimator in that the matrix \hat{G}_W is used in place of $(X'WX)\sigma^2$.

A multiple R -squared statistic is computed for models with an intercept. An F -test for the overall regression is always computed and an option for testing subsets of coefficients is provided.

4.4 Logistic Regression

Estimates of the multivariate logistic model are obtained with this option. The algorithms for logistic regression were developed after the initial version of PC CARP was completed. Because the mean function for the logistic model is nonlinear in the parameters, the estimates are computed using an iterative weighted least squares algorithm. The variances of the estimates are computed by the extension to nonlinear estimation of the procedures given in Fuller (1975). See also Binder (1983). The basic operation of the Logistic Regression option is the same as that of the Regression option. For example, independent and dependent variables are specified in the same way.

4.5 Post Stratification

After completion of the original PC CARP program a supplement for post stratification was developed for many of the estimators. The post stratification is assumed to be that in which the weights have been adjusted to produce estimates for certain categories that match known population totals. This type of post stratification is called gamma post stratification by Fuller and Sullivan (1987).

The program computes the variance of the post stratification estimator based on a representation in which the estimator is expressed as a sum of ratio estimators.

4.6 Stratum Collapse

For purposes of variance computation, the user may use the collapse option to eliminate one cluster strata. If this option is chosen, every one-cluster stratum is grouped with the immediately following stratum in the data set. The stratum and cluster identification of the involved records are changed to reflect the new stratification. If stratum sampling rates are present, new rates are defined by

$$f_i^* = (n_i f_i^{-1} + n_{i+1} f_{i+1}^{-1})^{-1} (n_i + n_{i+1}),$$

where stratum i , with $n_i = 1$, has been combined with stratum $i + 1$. These new rates are also saved in an auxiliary rate file for possible future use. Different orderings of the strata will produce

different collapsed data sets and different collapsed stratum rates. The program requires an additional pass through the data when either the collapse or the two-stage option is selected.

4.7 Hot Deck Imputation

PC CARP requires a complete data set for analysis. Many practitioners will write a special program, or use one of the readily available PC programs to edit their data and to impute for missing values.

For those desiring it, a hot deck imputation program, called PRE CARP, is provided with PC CARP. The hot deck operation replaces a missing value with the value for the same item from the record immediately preceeding the missing record in the data file. PRE CARP permits the user to specify a classification variable, containing up to ten categories, such that the missing value is replaced by the preceeding record in the same category. PRE CARP will also create an indicator variable for each variable with missing values. This indicator variable can then be used with the subpopulation option to compute means based on the original observations.

5. EXAMPLES

In this section, several analyses are performed with a constructed data set and run times are presented. The purpose of the test runs is not to examine all possible combinations of factors influencing processing time, but rather to give an idea of the time required to run some of the available program analyses.

The test data were constructed from a subset of the second National Health and Nutrition Examination Survey (NHANES II). The test data set has 2400 observations which are divided into 32 strata. Each stratum has two primary sampling units and the primary sampling units are of varying sizes. Each observation also has a non-zero sampling weight.

Figure 1. Output for Example C, Mean Age by Sex and Race Combinations

Subpopulation Means				
Dependent variable is Age				
Category	Estimate	S.E.	C.V.	DEFF
Sex = 1.0000	Race = 1.0000			
	3.06811D+01	6.19678D-01	2.0197D-02	1.3967D+00
Sex = 1.0000	Race = 2.0000			
	3.13016D+01	7.88580D-01	2.5193D-02	9.5384D-01
Sex = 1.0000	Race = 3.0000			
	3.41579D+01	2.24111D+00	6.5610D-02	2.0965D+00
Sex = 2.0000	Race = 1.0000			
	1.33742D+01	3.18904D-01	2.3845D-02	1.2588D+00
Sex = 2.0000	Race = 2.0000			

Sex = 2.0000	Race = 3.0000			
	1.71957D+01	9.53816D-01	5.5468D-02	1.1389D+00

Figure 2. Univariate Output for Nonfarm Households for Example D

UNIVARIATE 1				
Classification variable is Farm and its level is 1				
Number of Sample Elements in Subpopulation = 111				
Dependent variable is Age				
Subpopulation Variance = 4.25645D+02				
Subpopulation C.V. = 7.14101D-01				
Subpopulation Mean				
Estimate	S.E.	C.V.	DEFF	
2.8891089D+01	2.2543875D+00	7.80305D-02	9.86336D-01	
Extreme Values of Sample Elements in Subpopulation				
Smallest	Number of		First Observation ID	
Values	Observed Values	Stratum	Cluster	Weight
1.000D+00	1	32	1	3.000D+00
2.000D+00	1	15	1	2.000D+00
3.000D+00	3	10	2	3.000D+00
Largest	Number of		First Observation ID	
Values	Observed Values	Stratum	Cluster	Weight
7.400D+01	2	29	1	3.000D+00
7.100D+01	2	7	1	2.000D+00
7.000D+01	4	7	1	2.000D+00
Quantiles				
	Estimate	S.E.	95% Confidence Interval	
0.01	2.2690811D+00	7.3167585D-01	(8.05729D-01, 3.73243D+00)	
0.05	4.2364814D+00	1.1977759D+00	(3.34942D+00, 8.14053D+00)	
0.10	7.7750203D+00	1.3563759D+00	(5.36691D+00, 1.07924D+01)	
0.25	1.3652930D+01	1.4225576D+00	(9.99238D+00, 1.56826D+01)	
0.50	1.9449315D+01	2.2740912D+00	(1.57192D+01, 2.48156D+01)	
0.75	4.5698071D+01	4.7577709D+00	(3.58936D+01, 5.49247D+01)	
0.90	6.2787426D+01	2.3472775D+00	(5.51417D+01, 6.45308D+01)	
0.95	6.5923423D+01	1.2228344D+00	(6.43837D+01, 6.92750D+01)	
0.99	7.1714993D+01	1.1425033D+00	(7.05401D+01, 7.40000D+01)	
Interquartile Range				
	Estimate	S.E.		
	3.2045141D+01	4.2434890D+00		

The variables in the data set are:

- 1. Sex 1 = male, 2 = female
- 2. Race 1 = white, 2 = black, 3 = other
- 3. Farm 1 = non-farm household, 2 = farm household
- 4. Income Household income in thousands of dollars
- 5. Age Age in years.

A variable whose value is one for every observation (intercept variable) was created by the program. The analyses performed were:

- A. Mean income for the sampled population
- B. Mean income by stratum
- C. Mean age for the two way classification of sex and race
- D. Sample distribution functions of age for farm and non-farm groups.

Analysis A, estimating mean income, was performed using the Ratio option with Income as the numerator variable and the intercept variable in the denominator. The estimates of mean

income by stratum, Analysis B, were computed directly with the Stratum Means option. Analysis C was performed with the Subpopulation Means option by crossing the classification variables Sex and Race and specifying Age as the dependent variable. The output from this analysis is given in Figure 1. The symbols "*****" under the classification "Sex = 2 Race = 2" indicates that there were no observations falling into that classification category. The values of the design effects underscore the importance of taking into account the sampling design in the computation of estimated variances. For example, the design effect for the estimate with Sex = 1 and Race = 3 is approximately two. This means that the estimated variance of the sample mean for a simple random sample is one half of the variance estimate for the stratified cluster sampling plan. Characteristics of the distribution of Age for each of the two levels of the variable "Farm" were estimated using the Univariate option. The portion of the output for this example that pertains to nonfarm households is given in Figure 2. All the variances and standard error estimates given in this output take into account the sampling design.

The run times (in seconds) for analyses A, B, C and D for the 2,400 observations were 70, 135, 120 and 360, respectively. The runs were made on an IBM PC AT with the data stored on the hard disk and read in free format. Stratum sampling rates were not entered into the program. Output was routed to the monitor and to a disk file. Design effects for the estimates were requested in all of the analyses. The first three analyses require only one pass through the data for each analysis. More statistics are computed for analyses B and C than for analysis A. Analysis D requires 4 passes through the data, two passes for each univariate analysis.

ACKNOWLEDGEMENTS

In addition to the U.S. Census Bureau, support for PC CARP was provided by the Soil Conservation Service, U.S. Department of Agriculture. We thank the referees and the editors for useful comments. Persons in developing countries who are interested in the PC CARP program should contact: PC CARP, International Statistical Programs Center, U.S. Bureau of the Census, Washington, D.C., USA. 20233. Other interested persons should contact: PC CARP, Statistical Laboratory, Iowa State University, Ames, Iowa, USA 50011. Copies have been made available to institutions in more than 20 countries.

REFERENCES

- BARR, A.J., GOODNIGHT, J.H., SALL, J.P., BLAIR, W.H., and CHILKO, D.M. (1979). SAS User's Guide. SAS Institute, Raleigh, North Carolina.
- BINDER, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.
- COCHRAN, W.G. (1977). *Sampling Techniques*. New York: John Wiley.
- DISKIN, B.A. (1985). Microcomputers in developing country statistical offices: current use and a look to the future. Proceedings of the 45th Session of the International Statistical Institute, Amsterdam.
- FRANCISCO, C.A. (1987). Estimation of quantiles and the interquartile range in complex surveys. Unpublished Ph.D. dissertation. Iowa State University, Ames, Iowa.
- FRANKEL, M.R. (1971). *Inference from Survey Samples*. Institute of Social Research, University of Michigan, Ann Arbor.
- FULLER, W.A. (1975). Regression analysis for sample survey. *Sankhyā C* 37, 117-132.
- FULLER, W.A., and SULLIVAN, G. (1987). Gamma Post Stratification. Report to the U.S. Bureau of the Census. Dept. of Statistics, Iowa State University, Ames, Iowa.

- HERRAMAN, C. (1968). Sums of squares and products matrix. *Applied Statistics*, 17, 289-292.
- HIDIROGLOU, M.A. (1974). Estimation of regression parameters for finite populations. Unpublished Ph.D. thesis, Iowa State University, Ames, Iowa.
- HIDIROGLOU, M.A., FULLER, W.A., and HICKMAN, R.D. (1980). *SUPER CARP*, Dept. of Statistics, Iowa State University, Ames, Iowa.
- KISH, L. (1965). *Survey Sampling*. New York: John Wiley.
- LONGLEY, J.W. (1967). An appraisal of least squares programs for the electronic computer from the point of view of the user. *Journal of the American Statistical Association*, 62, 819-841.
- PARK, H.J. (1987). Univariate Analysis in PC CARP. Unpublished Creative Component for the M.S., Iowa State University, Ames, Iowa.
- RAO, J.N.K., and SCOTT, A.J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *Annals of Statistics* 12, 46-60.
- WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

The 1986 Test of Adjustment Related Operations in Central Los Angeles County

GREGG DIFFENDAL¹

ABSTRACT

As part of the planning for the 1990 Decennial Census, the Census Bureau investigated the feasibility of adjusting the census for the estimated undercount. A test census was conducted in Central Los Angeles County, in a mostly Hispanic area, in order to test the timing and operational aspects of adjusting the Census using a post-enumeration survey (PES). This paper presents the methodology and the results in producing a census that is adjusted for the population missed by the enumeration. The methodology used to adjust the test census included the sample design, dual-system estimation and small area estimation. The sample design used a block sample with blocks stratified by race/ethnicity. Matching was done by the computer with clerical review and resolution. The dual-system estimator, also called the Petersen estimator or capture-recapture, was used to estimate the population. Because of the nature of the census enumeration, corrections were made to the census counts before using them in the dual-system estimator. Before adjusting the small areas, a regression model was fit to the adjustment factor (the dual-system estimate divided by the census count) to reduce the effects of sampling variability. A synthetic estimator was used to carry the adjustment down to the block level. The results of the dual-system estimates are presented for the test site by the three major race/ethnic groups (Hispanic, Asian, Other) by tenure, by age and by sex. Summaries of the small area adjustments of the census enumeration, by block, are presented and discussed.

KEY WORDS: Census undercount; Dual-system estimation; Synthetic estimation; Post-enumeration survey.

1. INTRODUCTION

Since the first U.S. Census in 1790, problems have existed in finding and counting every person who should be counted. Advances in demographics and statistics have permitted census coverage estimates to be produced, beginning with the 1950 census. Coverage estimates have been used to evaluate census shortfalls and determine areas of needed improvements for succeeding censuses. The census coverage estimates have shown a steady improvement in census taking since the 1950 estimates were produced. One series of estimates shows the U.S. level undercount was 4.4% for 1950, 3.3% for 1960, 2.8% for 1970 and 1% for 1980. Despite this continuing reduction in the percent undercount, estimates remain higher for certain groups in the U.S. For example, the black undercount has remained about 5 percent above the national average.

Results also indicate high undercounts are measured for other ethnic groups-especially the Hispanic population. Central cities have higher undercounts as do rural areas. Males have higher undercounts than females. The age group 20 to 45 also has a high undercount.

The methods used since 1950 to measure the undercount in the U.S. are a post-enumeration survey (PES) and demographic analysis. The Census Bureau has announced that these will be the major tools to estimate the undercount for the 1990 census. A PES uses an independent sample of persons that are matched to the census to estimate the total population. Marks (1978)

¹ Gregg Diffendal, Undercount Research Staff, Statistical Research Division, Bureau of the Census, Washington, D.C. 20233, USA. This paper reports research undertaken by a member of the Census Bureau's staff. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau.

and U.S. Census Bureau (1979) describe previous work on using a PES to measure census coverage. Demographic analysis uses birth, death and other administrative records to estimate the total population in the U.S. Fay *et al.* (1988) describe the 1980 undercount estimates from demographic analysis and the Post-Enumeration Program (PEP).

In 1980, increased scrutiny of census numbers resulted in a number of court suits arguing for an adjustment of the 1980 census counts. Some of the issues that led to the court suits include: the existence of the differential undercount between blacks and nonblacks; the introduction of revenue sharing in the 1970's which tied monies directly to population counts; and declining populations in some cities and states which have traditionally had higher undercounts. The U.S. Census Bureau argued against adjustment of the 1980 census for the measured undercount on the basis that the measurement was error prone and an adjustment would not improve the unadjusted census counts.

The Census Bureau did embark on a research program after 1980 to evaluate alternate methods and ways to improve the undercount measurement process (Mulry *et al.* 1981 and Hogan 1984). Hogan (1984) proposed a series of tests to improve the undercount measurements in conjunction with the test censuses. These started with a PES in Tampa, Florida in 1985 to test and evaluate computer matching. This test verified the feasibility of computer matching (Jaro and Childers 1986). Test censuses and PES's were also conducted in 1986 in Los Angeles and Mississippi. A PES was conducted in Los Angeles to test the timing and operational aspects of adjusting the census. In Mississippi, a PES was conducted to evaluate the PES operations in a rural test site (Anolik 1988).

A Pre-Enumeration Survey was also conducted in 1986 in Los Angeles to determine if further gains in timing could be obtained if some of the field work was conducted before the census rather than after the census enumeration as in a PES (Wolfgang 1987). A PES was conducted in 1987 in rural North Dakota for evaluation of the PES operations in rural areas, where a door-to-door enumeration is conducted rather than a mail-out census as in the other test sites. Finally, work is under way for the 1988 Census Dress Rehearsal. The Dress Rehearsal will be used to test all census operations before conducting the 1990 Decennial Census.

The focus of this paper is on the 1986 PES in Central Los Angeles County, called the Test of Adjustment Related Operations (TARO), conducted in conjunction with the test census. The test site comprised three major race/ethnic groups: Hispanic, about 75% of the total population; Asian, about 15%, and Other, mostly white, with about 10% of the total population. The results of the PES show an estimated undercount of 9%. For the major race/ethnic groups in the test site, the Hispanic, Asian, and Other undercounts are estimated at 9.8%, 7.3%, and 6.2%, respectively. This paper describes the methodology and operational aspects of estimating these undercounts.

Section 2 presents the methodology used in 1986 to measure the undercount and how to incorporate the undercount estimates into the census count to produce an adjusted census. Section 3 discusses the schedule of operations in carrying out TARO including field operations and matching. Section 4 presents a summary of the undercount estimates for the poststrata and undercount estimates at the block level. Section 5 summarizes the major findings and presents some conclusions.

2. METHODOLOGY

2.1 Overview of Samples Used in Estimation

To estimate the population, the PES used two samples, called the P (for Population) sample and the E (for Enumeration) sample. The P sample is used to measure census omissions. The E sample is used to measure census erroneous enumerations.

The P sample consists of a block sample with an an independent listing of housing units and personal interviews whereas E-sample data are the census enumerations (counts) from the same sample block. The P sample obtained data needed for matching and estimation including census day residence. A design decision was made that defined who is included in the P sample. The P sample was all persons living at the sampled address at the time of the PES interview. The alternate procedure would interview the residents on census day. We decided against the latter approach because all movers involve proxy respondents (interview is from nonhousehold members). For the approach chosen, movers were living at the sample address and can have completed interviews without resorting to a proxy respondent. However all residents on census day who moved outside the test site before the PES interview have zero probability of being captured in the P Sample. All P-sample persons who lived outside the test site on census day were considered out-of-scope.

After interviewing, all P-sample persons were matched to the census. A computer matching program was used with clerical review. A second design decision defined the extent of search for matching. The PES classified a P-sample person as matched if the person was counted in the census anywhere in the test site. An alternate procedure would define a more limited search area, such as the PES block and neighboring blocks. Then a P-sample person is called a match only if the corresponding census person is within this search area. As an aside, the 1990 PES procedure will use a limited search area for matching.

All unresolved cases from matching were sent to followup to obtain additional information for matching. The followup workload from the P sample was greatly reduced by asking all questions needed for matching at the time of the original interview. Therefore only incomplete personal characteristics, incomplete mover address, and uncertain match cases were sent to followup from the P sample. Nonmatched P-sample cases were considered resolved and not sent to followup. Many E-sample persons are matched to P-sample persons and are resolved without the need of another interview. All E-sample persons not resolved from the P-sample interview were sent out for a followup interview that is used to determine their enumeration status. Operational aspects are discussed in more detail in the following section. The types of census erroneous enumerations measured by the E sample included geocoding error, duplication, fabrication, persons born after census day, persons who died before census day and unmatchable cases. Geocoding error is defined as a census enumeration that exist outside the search area, the entire test site. Unmatchable cases are census enumeration without a name. Unmatchable cases cause an overestimate of the number of erroneous enumeration, but are treated in a similar manner as erroneous enumerations in the estimator.

2.2 Dual-System Estimation

In order to estimate the total population, a dual-system estimator is used which combines the information from the P and E samples. Wolter (1986) describes different dual-system estimators and their underlying assumptions. The dual-system estimator used in TARO is written

$$DSE = \frac{N_p(CEN-SUB-EE)}{M}, \tag{1}$$

where N_p = estimator of the total PES population, CEN = unadjusted census count, SUB = number of census whole-person substitutions, EE = estimator of the number of erroneous enumeration and unmatchable persons included in the census, derived from the E sample, M = estimator of the number of persons in both the census and the PES populations. Census whole-person substitutions are defined as any person included in the census with fewer

Table 1
Dual-System Classification

		P-Sample Target Population		
		In	Out	Total
Census	In	N_{11}	N_{12}	N_{1+}
Enumeration	Out	N_{21}	N_{22}	N_{2+}
	Total	N_{+1}	N_{+2}	N_{++}

than two demographic characteristic. In order to better understand and explain some of the unique features of the dual-system estimator, Table 1 shows the classification of each person in the population.

The population quantities in Table 1 are estimated by components of the dual-system estimators: $\hat{N}_{11} = M$, $\hat{N}_{+1} = N_p$, $\hat{N}_{1+} = \text{CEN-SUB-EE}$. The value of N_{22} is unobservable by definition but is estimated by assuming independence between the census enumeration and the P sample of the PES. The estimate of N_{22} is given by

$$\hat{N}_{22} = \hat{N}_{12}\hat{N}_{21}/\hat{N}_{11}. \tag{2}$$

By using the estimators defined above, the estimate of the total population is given by $\hat{N}_{++} = \text{DSE}$.

Because of problems in matching census data, special handling is needed to prevent an overestimate of the population. The dual-system estimator assumes every person is uniquely assigned to one cell in Table 1. So instead of just using the census count, the estimate of erroneous enumerations is subtracted from the census count to give an estimate of the number of unique persons counted in the census. Additionally, the dual-system estimator assumes each person can be called a match or a nonmatch. Census enumerations with insufficient information for matching (e.g., no name or fewer than two demographic characteristics) cannot be called matches or nonmatches with certainty. Therefore, unmatchable persons are also subtracted from the census count. All corresponding P-sample persons are called nonmatches and assigned to the N_{21} cell.

2.3 Sample Design

The sample design was a stratified sample with the sampling unit being a block. Two types of data were used to stratify the test site — a count of housing units by block obtained from the 1986 census address file and a mapping of 1980 census race data into the 1986 census geographic units. This mapping could only be made at the census tract level which equals one to six blocks. Therefore, the assignment of the racial grouping was done at the census tract level. All blocks within the census tract were assigned to the same racial category, and thus were in the same stratum.

The test site was stratified into six sampling strata, described in Table 2.

All blocks with special places (mostly group quarter population) were put into a separate sampling stratum. These blocks were considered out-of-scope and were not sampled. Small blocks were placed in a separate stratum to reduce the sampling variance. All blocks in census tracts with at least 18% Asian defined the Asian strata. All non-Asian blocks in census tracts with at least 40% Hispanic defined the three Hispanic strata. All remaining blocks that were not in the above strata defined the Other strata.

Table 2
Sampling Strata and Allocation of Sampled Blocks

Sampling Strata	Number of Blocks Sampled
1. Hispanic Blocks with large multiunits	8
2. Hispanic Blocks with small multiunits	49
3. Hispanic Blocks with single units	39
4. Asian Blocks	35
5. Other Blocks	38
6. Blocks with two or fewer housing units	21

The 1986 housing count data also contained information on single unit and multiunit structures. These data were used to split the Hispanic strata into single unit, small multiunit, and large multiunits. The Hispanic large multiunits stratum was defined as the Hispanic blocks with 50% or more of the housing units in structures with 10 or more addresses. The Hispanic single unit stratum was defined as the Hispanic block with more than 50% of the housing units in single units. The Hispanic small multiunits stratum was defined as the remainder of the Hispanic blocks.

Within each of the sampling strata, an equal probability systematic sample of blocks was chosen. The sample consisted of 190 blocks containing about 6000 housing units. Table 2 contains the breakdown of the sampled blocks by the sampling strata. Large blocks with 70 or more housing units were subsampled to reduce the interviewing workload. The subsampling consisted of splitting the block into clusters of 35 to 50 housing units, using address ranges or block faces. One cluster was randomly selected for P-sample interviewing. The E sample was defined as all persons the census counted in the same cluster.

2.4 Poststratification

The dual-system estimator is biased and the bias can be large if the undercount rates are significantly different for subgroups of the population (Wolter 1986). To control this bias, the test site was partitioned into groups (poststrata) felt to have the similar undercount rates. Dual-system estimates were then calculated within each poststratum.

The poststrata were chosen by examining the test site composition and from analysis of the 1980 PES data. The most important discriminating variable of the undercount was race. Three race-ethnic groups were used: Hispanic, Asian and Other. A separate poststratum for blacks was not possible since few blacks lived in the test site. Minority renter was an important explanatory variable in our previous research (Isaki *et al.* 1987). Therefore, tenure was also used in constructing the poststrata. Hispanics living in a block with fewer than 50% of the population being Hispanic (called Non-Hispanic blocks) were thought to have a different undercount rate from other Hispanics and was assigned to a separate poststratum. Table 3 shows the seven race-tenure groups which are crossed by age (0-14, 15-29, 30-44, 45-64, 65 +) and sex to give the 70 poststrata used in estimation.

Table 3 also shows the sample sizes for the P sample and for the E sample. The lower sample size for the P sample than the E sample is partly explained by in-movers in the P sample which are treated as being out-of-scope.

Table 3
Race-Tenure Categories Used in Poststratification,
Including Sample Sizes

Race-Tenure Categories	P Sample	E Sample
Hispanic Renters in Hispanic Blocks	8,182	8,739
Hispanic Owners in Hispanic Blocks	5,688	5,867
Hispanics in Non-Hispanic Blocks	896	1,005
Asian Renters	666	911
Asian Owners	1,144	1,230
Other Renters	1,135	1,316
Other Owners	1,841	1,908
Total	19,552	20,976

2.5 Handling Missing Data

To compute the dual-system estimates, a complete data file is needed. The 1986 test contained missing data, as is true for any sample survey. Schenker (1988) presents a description of the methods used to handle missing data, including some effects of different assumptions about missing data on the dual-system estimates. For completeness, we give a brief description of the methods.

Missing data occurred for person and household characteristics, the match status (matched/nonmatched) for the P-sample persons, and enumeration status (correct/erroneous) for the E-sample persons. For P-sample noninterviews, a weighting adjustment was used. Missing characteristics were imputed using a “hot-deck” procedure. For match status, a logistic regression model was used to estimate the probability of being matched. Rather than assign a definite match or nonmatch status to each unresolved case, the estimated probabilities were used in the dual-system estimates. An analogous procedure was used for missing E-sample enumeration statuses.

2.6 Small Area Estimation

To make an adjustment additive at all levels of aggregation for users, the estimates of the undercount are carried down to the block level (the smallest geographical unit). But before carrying the undercount estimates to the block level, a regression model is used to “smooth” the effects of sampling error. Adjustment factors are used as the dependent variable in the regression model. An adjustment factor is defined as the dual-system estimator divided by the census count:

$$Y = \text{DSE}/\text{CEN}, \tag{3}$$

where CEN and DSE were defined previously.

The regression model is written as

$$Y_i = B_0 + B_1X_{i1} + \dots + B_pX_{ip} + S_i + E_i, \tag{4}$$

where Y_i = adjustment factor for the i -th poststratum ($i = 1, \dots, 70$), X_{ij} = independent variable ($j = 1, \dots, p$), B_j = regression coefficient to be estimated, S_i = sampling error of the adjustment factor, E_i = model error, and the S_i and E_i are independent and normally distributed with mean 0 and variances equal to σ_i^2 and ϵ^2 respectively. The ϵ^2 and B_j 's are estimated using maximum likelihood methods (Ericksen and Kadane 1985). The σ_i^2 are estimated directly from the sample. The sample-based adjustment factor and the model-based adjustment factor are averaged together to form the predicted adjustment factor

$$AF_i = \left(Y_i/\sigma_i^2 + \sum_j X_{ij}\hat{B}_j/\epsilon^2 \right) \left(\sigma_i^{-2} + \epsilon^{-2} \right)^{-1}, \tag{5}$$

which is used to adjust the census block data. The variance of AF_i can be obtained from the results in Freedman and Navidi (1986).

Synthetic estimation was used to carry down the adjustment from each poststratum to the census block. The synthetic estimator is written as

$$ADJ_{ij} = AF_i \times CEN_{ij}, \tag{6}$$

where i and j denote the poststratum and block respectively and ADJ is the adjusted population at the block level.

The adjusted block population, ADJ_{ij} , is usually a noninteger number. The census counts whole persons. In order to incorporate the adjustment into the census, the noninteger values must be transformed into integers. Integerization (or controlled rounding) rounds all values to the integer part of the number or to the integer part of the number plus one (Causey *et al.* 1985).

After integerization of the adjusted block estimates, counts were produced for the number of persons by age-race-sex to be added to or subtracted from each block. In the case of undercounts, a census enumeration having the same range of characteristics as the estimated missed person was randomly selected from within the block and copied into a new census record. A nonhousehold category was used to add persons to the census so that household relationships and creation of new households were not needed. Zaslavsky (1988) describes an alternate procedure, using weighting, for adding persons and households to census blocks. In the case of overcounts, census persons with the required characteristics would be flagged and would not be counted in the adjusted census tabulations.

3. OPERATIONS AND TIMING

The major focus of this test census was to study the timing and operational aspects of adjusting the census. Previous PES's at the Census Bureau have taken about two years or longer to complete. For example, the 1980 PES produced undercount estimates in the fall of 1981 and a final set of estimates in early 1982.

Table 4 presents the major census and PES operations and their start and end dates. Gaps exist in Table 4 because all census and PES operations are not listed. Some PES operations have overlapping time schedules since these operations were occurring at the same time. PES activities started after all major census field activities were completed. This helps ensure independence between the census and the PES by having the field staffs working at different times.

Table 4
1986 TARO Operational Schedule

Operation	Start	End
Census Day	March 16	March 16
Nonresponse Followup	April 09	May 08
Key Census Names	May 23	June 10
Census File for Matching	Aug. 08	Aug. 15
PES Address Listing	June 17	June 21
PES Subsampling	June 25	July 01
PES Interviewing	June 25	Aug. 08
Key PES form	July 21	Aug. 19
Computer Match	Aug. 28	Sept. 09
Extended Computer Match	Sept. 09	Oct. 03
Clerical Match	Sept. 15	Oct. 31
Field Followup	Sept. 23	Nov. 06
Followup Matching	Sept. 29	Nov. 06
Key Match Results	Oct. 21	Nov. 10
Prepare P- and E- sample files	Nov. 11	Jan. 02
Imputations	Jan. 05	Jan. 11
Final Census file	-	Jan. 05
Estimate Poststrata	Jan. 12	Feb. 11
Small Area Estimates	Feb. 12	Feb. 22

The census was conducted by mailing a questionnaire to every known housing unit and asking a household member to complete the form on Census Day (March 16). Each household that failed to mail back its questionnaire was completed in person by an enumerator. This is called nonresponse followup. Completed forms were sent to the processing office for entering the data, which included for this test all census names, into the computer.

The first step of the PES produced an independent listing of all addresses in the sample blocks. The listings were compared to an administrative list to ensure accuracy and completeness. This quality control check showed that 127 (67%) blocks had no change to the address listing. The remaining 63 (33%) blocks had changes made from the quality control check and were relisted. The relisting added addresses to 37 blocks, corrected addresses in 39 blocks, and deleted addresses in 9 blocks. (Since multiple changes were made for some blocks, the above numbers do not add up to the total number of relisted blocks.) The changes in the address listings from the quality control check showed only minor corrections. After passing the quality control check, all blocks of 70 or more housing units were subsampled using block faces or address ranges.

The PES interview was conducted by personal visits. Questions were asked of all current residents to obtain their demographic characteristics. Special questions asked about residence on Census Day, mailing address, alternate addresses such as college residence, and other persons who may have lived at this residence on Census Day. A quality control check of the PES questionnaire verified the roster of names. For the sample of forms checked, 96% passed the quality control operation. The 4% that failed the quality control check were reinterviewed and corrected.

The final outcome of the interviewing showed that 5,714 (93.2%) of the housing units had a completed interview with a household member. Another 193 (3.1%) housing units were vacant and 189 (3.1%) housing units had a completed interview with a non-household member (e.g. neighbor). Only 32 (0.5%) housing units were coded as noninterviews. The extremely low noninterview rate is attributable to the 5 week interviewing period.

As the PES questionnaires were completed they were prepared for computer matching to the census file. The computer matching was split into two parts: first, matching the PES data with the E-sample data and second, called extended computer matching, matching all P-sample cases that did not match in the first part of the computer matching to the remaining census data. The extended computer match was used to match movers between Census Day and the time of the PES interview and geographical coding errors, i.e., where the housing unit is assigned to the wrong block. The first part of the computer matching assigned a match to 14,700 (73.5%) of the P-sample cases and assigned a possible match to another 2,550 (12.0%). The extended computer matching assigned a match to another 130 persons (0.7%) and assigned a possible match to another 570 persons (2.9%). Because the extended computer matching assigned a match status to only a small percentage of P-sample cases, we concluded that the geographical coding in Los Angeles had few errors.

Clerical matching reviewed the results of the computer matching. Clerical matching also identified the cases with insufficient data for matching (for which imputation is necessary). Clerical matching prepared followup forms for unresolved P-sample and E-sample cases.

Field followup consisted of 1,551 housing units with 1,511 (97.4%) being recorded as completed interviews. The field followup was followed by final matching. The final P-sample results show that 17,018 (85.2%) persons were matched to a census persons and 2,373 (11.9%) persons were not matched. Another 426 (2.1%) persons were considered out-of-scope (mostly persons who lived outside the test site on Census Day) and 161 (0.8%) persons were unresolved (and later had match status imputed). The final E-sample results show that 19,637 (93.6%) persons were correctly enumerated and 360 (1.7%) were erroneously enumerated in the census. Another 976 (4.7%) persons were unresolved and had an enumeration status imputed.

All missing data after final matching including match status for the P sample and enumeration status for the E sample were imputed. The results were used to create the dual-system estimates. The estimates were smoothed and carried down to the block level to create an adjusted census file. The improvements in timing to produce the undercount estimates were mainly due to the matching activities. The computer and clerical matching for TARO took about 3 months, while the 1980 PEP matching activities took over one year to complete. Additional time savings were due to improved planning of operations and better access of census materials.

4. ESTIMATES

4.1 Poststrata Estimates

This section presents the undercount estimates for various aggregations of the poststrata. Table 5 presents the percent undercount $100(1-CEN/DSE)$, percent nonmatched $100(1-M/N_p)$, percent erroneously enumerated $100(EE/CEN)$, and percent substituted $100(SUB/CEN)$.

A feature of the dual-system estimator is that the estimates summed over several categories does not equal the direct estimate of the summed categories. To keep the estimates reported in Table 5 consistent, all estimates are summed over the other relevant categories.

Examining Table 5 for percent undercount by the race-tenure groups, one concludes: tenure is a good stratification variable with higher undercount estimates for renters than for owners; race/ethnicity also differentiates the undercount with higher undercount estimates for Hispanics than for Asians, which in turn are higher than for Others. Percent erroneously enumerated is higher for renters than for owners, but almost no differences between the race/ethnicity groups. Percent substituted is higher for Hispanic and Other renters than for Hispanic and Other owners. Asian owners had a higher percent substituted than Asian renters, the reverse from the two other race-ethnicity groups.

Table 5
Percent Undercount and the Components of the Dual-System Estimates for the Poststrata

Post-Strata	Percent Undercount ^a	Percent Nonmatched of the P-Sample	Percent Erroneously Enumerated of the E-sample ^b	Percent Substituted of the Census
Hispanic Renters in Hispanic Blocks	13.7	17.1	2.6	1.7
Hispanic Owners in Hispanic Blocks	5.5	8.1	1.2	1.5
Hispanics in Non-Hispanic Blocks	7.5	9.7	1.4	1.3
Asian Renters	11.1	13.4	2.1	1.2
Asian Owners	4.6	6.8	1.2	1.5
Other Renters	9.9	12.9	2.4	1.7
Other Owners	3.8	5.8	1.3	0.9
0-14	8.8	11.9	2.2	1.6
15-29	13.6	16.2	2.1	1.6
30-44	8.6	10.8	1.4	1.4
45-64	4.5	6.6	1.3	1.4
65 +	3.3	5.9	1.7	1.3
Male	9.7	12.1	1.7	1.5
Female	8.3	10.8	1.9	1.5
Total	9.0	11.4	1.8	1.5

^a All estimates are summed over all other categories.

^b Erroneously enumerated includes unmatchable nonsubstituted.

Examining Table 5 for percent undercount by age and sex one observes that the age group 15-29 had the highest undercount and males have a higher undercount than females. The age groups 0-14 and 30-44 have similar undercount estimates, slightly below average for the test site. The age groups 45-64 and 65 + also have similar undercount estimates, well below the other age groups. These results are fairly consistent in distribution with previous undercount results.

Percent erroneously enumerated is highest for the two youngest age group 0- 14 and 15-29. The age groups 30-44 and 45-64 have similar low estimates of percent erroneous enumerated. Surprisingly the percent erroneously enumerated is in the middle for the age group 65 + . Small differences are observed in percent erroneous enumeration for the sex groups. Only small differences are observed for percent substituted for the age groups or the sex groups.

4.2 Small Area Estimates

Before applying the adjustment at the block level, as mentioned earlier, a regression model was fitted to "smooth" the data and reduce the effects of sampling variability. The regression

model was fit to the 70 adjustment factors as defined by the poststrata. The regression modelling is used to find a common pattern of undercounting in the data. Then the sample-estimated adjustment factors are shrunk toward this common pattern. This is similar in spirit to the James-Stein estimator and empirical Bayes estimators. The independent variables that were available to use in the model were indicator variables for the race-tenure groups, for the age groups, and for the sex groups. No interaction terms were allowed to enter the model. The model that fit the data and had significant coefficients (under an unweighted regression model) was the following:

$$\hat{Y} = 1.038 + .090(HR) + .044(AR) + .013(OR) + .058(A15-29) - .009(A45-64)$$

where \hat{Y} = model-based adjustment factor

HR = 1 if Hispanic Renter in Hispanic Blocks
= 0 otherwise

AR = 1 if Asian Renter in all Blocks
= 0 otherwise

OR = 1 if Other Renter in all Blocks
= 0 otherwise

A15-29 = 1 if age group 15-29
= 0 otherwise

A45-64 = 1 if age group 45-64
= 0 otherwise.

The regression model shows the larger undercount estimates for all renters over owners. Also the age group 15-29 has much higher undercount estimates than other age groups. The age group 45-64 has lower undercount estimates than the other age groups. The variable sex was statistically insignificant and was not included in the model. Two adjustment factors, Hispanics in Non-Hispanic blocks male 65 + and Asian renters male 65 + , had a zero estimated variance and were not included in the model. The predicted adjustment factor was defined as the sample-estimated adjustment factor for these two adjustment factors.

Table 6 contains the sample-estimated and predicted adjustment factors for the 70 poststrata. In general, the predicted adjustment factors lowers the highest estimated adjustment factors and raises the lowest estimated adjustment factors. The predicted adjustment factors have less variability than the sample-estimated adjustment factors. The most notable example of the effects of the regression model is for Asian renters female age 65 + . The predicted adjustment factor is 1.087 rather than the sample estimated adjustment factor of 1.212. This predicted adjustment factor is closer to the expectations of a lower undercount for the age group 65 + than for the other age groups.

The predicted adjustment factors were multiplied by the census counts for the 2,405 blocks in the test site. The adjusted census counts were rounded to form integer values. Although three predicted adjustment factors were less than one (an estimated overcount), the integerization process did not produce any adjusted overcounts.

The adjustment process added 32,843 people to the census, a 8.2% undercount rate. If the sample-estimated adjustment factors were used, then 36,454 people would have been added to the census, a 9.0% undercount rate. The process of smoothing lowered the undercount estimate by almost 10%. This occurred because the largest undercount estimates were lowered by the smoothing and these same groups had the largest population counts.

Table 6
Results of Smoothing TARO Adjustment Factors

Poststrata	Sex/Age	Estimated		Predicted	
		Adj. Factor Y	Std. Error	Adj. Factor AF	Std. Error
HR in HB	M 0-14	1.131	0.020	1.130	0.016
HR in HB	M 15-29	1.247	0.030	1.211	0.021
HR in HB	M 30-44	1.165	0.029	1.144	0.020
HR in HB	M 45-64	1.099	0.043	1.114	0.024
HR in HB	M 65+	1.055	0.044	1.110	0.023
HR in HB	F 0-14	1.124	0.023	1.126	0.018
HR in HB	F 15-29	1.234	0.032	1.203	0.022
HR in HB	F 30-44	1.084	0.017	1.098	0.015
HR in HB	F 45-64	1.125	0.040	1.121	0.024
HR in HB	F 65+	1.099	0.045	1.122	0.024
HO in HB	M 0-14	1.056	0.018	1.050	0.015
HO in HB	M 15-29	1.078	0.018	1.084	0.015
HO in HB	M 30-44	1.087	0.016	1.072	0.014
HO in HB	M 45-64	1.031	0.012	1.031	0.011
HO in HB	M 65+	1.073	0.028	1.054	0.019
HO in HB	F 0-14	1.059	0.020	1.051	0.016
HO in HB	F 15-29	1.088	0.016	1.090	0.014
HO in HB	F 30-44	1.033	0.012	1.034	0.011
HO in HB	F 45-64	1.020	0.012	1.022	0.011
HO in HB	F 65+	1.033	0.019	1.035	0.015
H in H'B	M 0-14	1.105	0.052	1.051	0.023
H in H'B	M 15-29	1.154	0.054	1.106	0.025
H in H'B	M 30-44	1.131	0.065	1.050	0.024
H in H'B	M 45-64	1.063	0.050	1.036	0.023
H in H'B	M 65+	0.991	0.000	0.991	0.000
H in H'B	F 0-14	1.137	0.047	1.059	0.023
H in H'B	F 15-29	1.033	0.022	1.060	0.017
H in H'B	F 30-44	1.079	0.037	1.051	0.021
H in H'B	F 45-64	1.033	0.028	1.031	0.019
H in H'B	F 65+	0.947	0.040	1.013	0.022
AR in all B	M 0-14	1.059	0.041	1.076	0.026
AR in all B	M 15-29	1.127	0.044	1.137	0.028
AR in all B	M 30-44	1.195	0.077	1.093	0.031
AR in all B	M 45-64	1.004	0.057	1.063	0.030
AR in all B	M 65+	0.982	0.000	0.982	0.000
AR in all B	F 0-14	1.067	0.047	1.079	0.028
AR in all B	F 15-29	1.215	0.055	1.153	0.029
AR in all B	F 30-44	1.173	0.105	1.087	0.032
AR in all B	F 45-64	1.012	0.061	1.065	0.030
AR in all B	F 65+	1.212	0.127	1.087	0.032

Table 6
Results of Smoothing TARO Adjustment Factors – Concluded

Poststrata	Sex/Age	Estimated		Predicted	
		Adj. Factor <i>Y</i>	Std. Error	Adj. Factor <i>AF</i>	Std. Error
AO in all B	M 0-14	1.045	0.030	1.041	0.019
AO in all B	M 15-29	1.059	0.038	1.085	0.022
AO in all B	M 30-44	1.091	0.040	1.053	0.022
AO in all B	M 45-64	1.035	0.020	1.033	0.016
AO in all B	M 65+	1.031	0.051	1.037	0.023
AO in all B	F 0-14	1.040	0.041	1.039	0.022
AO in all B	F 15-29	1.052	0.046	1.086	0.024
AO in all B	F 30-44	1.035	0.036	1.037	0.021
AO in all B	F 45-64	1.038	0.019	1.035	0.015
AO in all B	F 65+	1.051	0.045	1.041	0.022
O'R in all B	M 0-14	1.037	0.059	1.049	0.027
O'R in all B	M 15-29	1.252	0.114	1.115	0.031
O'R in all B	M 30-44	1.144	0.066	1.062	0.028
O'R in all B	M 45-64	1.055	0.031	1.047	0.022
O'R in all B	M 65+	1.068	0.056	1.054	0.027
O'R in all B	F 0-14	1.148	0.062	1.064	0.027
O'R in all B	F 15-29	1.126	0.054	1.112	0.028
O'R in all B	F 30-44	1.134	0.057	1.064	0.027
O'R in all B	F 45-64	1.068	0.041	1.049	0.025
O'R in all B	F 65+	0.948	0.021	0.992	0.018
O'O in all B	M 0-14	1.044	0.037	1.040	0.021
O'O in all B	M 15-29	1.148	0.064	1.103	0.025
O'O in all B	M 30-44	1.006	0.048	1.032	0.023
O'O in all B	M 45-64	1.036	0.017	1.034	0.014
O'O in all B	M 65+	1.017	0.019	1.025	0.016
O'O in all B	F 0-14	1.159	0.068	1.052	0.024
O'O in all B	F 15-29	1.081	0.042	1.092	0.023
O'O in all B	F 30-44	0.997	0.017	1.011	0.014
O'O in all B	F 45-64	1.025	0.012	1.026	0.011
O'O in all B	F 65+	0.997	0.012	1.004	0.011

Note: H: Hispanic, R: Renter, B: Block, M: Male, F: Female, O: Owner, O': Other, H': Non-Hispanic, A: Asian.
(Example: HR in HB: Hispanic Renter in Hispanic Block)

To summarize the block level adjustments, figure 1 shows the number of persons added by the number of blocks and figure 2 shows the percent of persons added by the number of blocks. Almost 80% of the blocks added less than 20 persons. Only 2 blocks added more than 150 persons. Those 2 blocks were fairly large, containing about 2,000 people each. Over 80% of the blocks had undercount estimates ranging from 4% to 12%. Many of the small blocks added a small percent of persons because the estimates were rounded down making a large change in the percent. The blocks with largest percent added were largely Hispanic and renters which had the largest predicted adjustment factors.

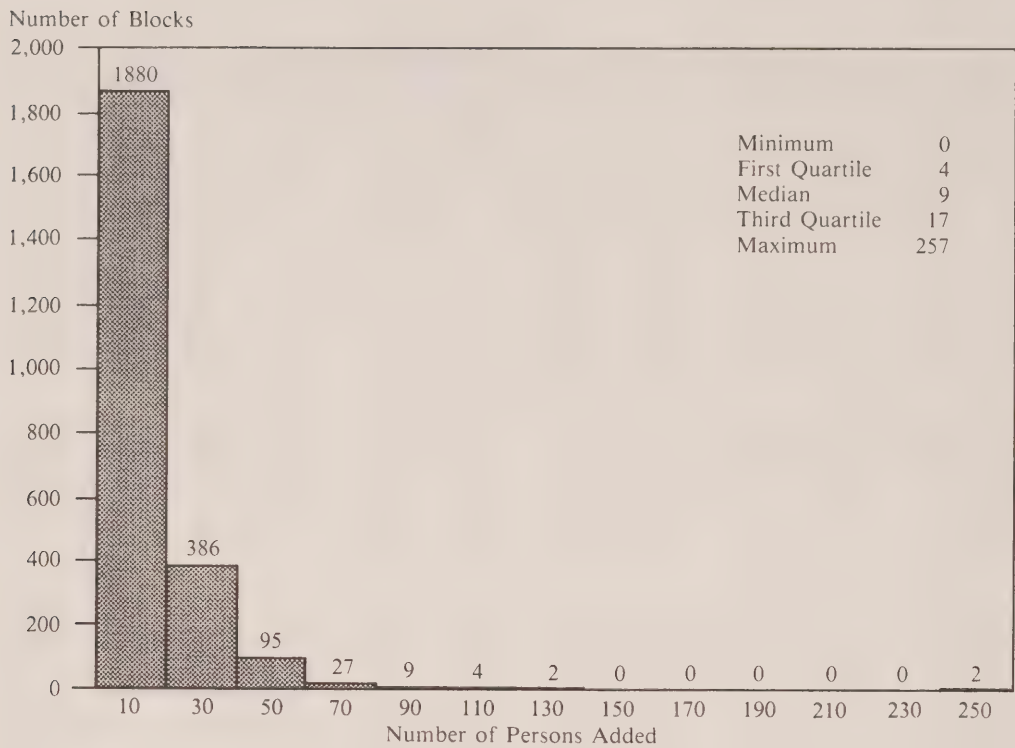


Figure 1. Number of Persons Added Per Block

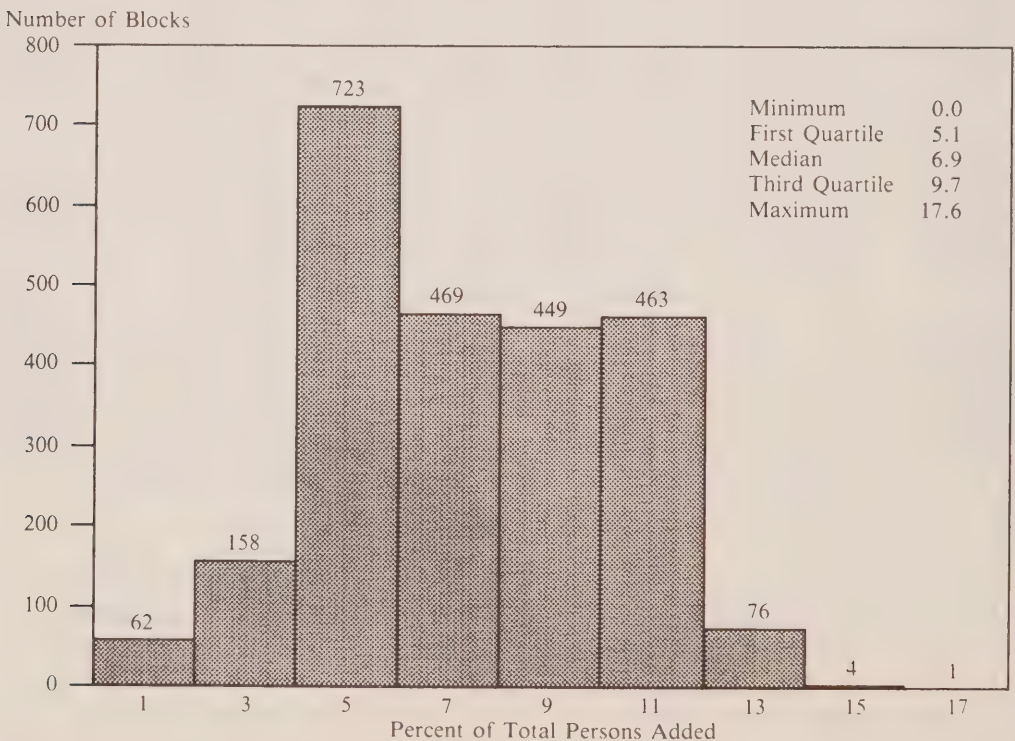


Figure 2. Percent of Total Persons Added Per Block

5. CONCLUSION

This paper discusses the methodology, operations, and the results of the Test of Adjustment Related Operations. TARO tested the operational and timing aspects of adjusting the census for estimated persons missed in the enumeration of the population.

The results from TARO demonstrate that undercount estimates can be produced in a timely manner. TARO was completed earlier than any previous PES.

TARO measured an undercount of 9% for the Central Los Angeles Count's test census. Separate dual-system estimates are presented for 70 race-tenure by age by sex categories. The dual-system estimates were smoothed by fitting a regression model to the estimates and then the resulting estimates were carried down to the block level. The use of block level undercount estimates allows aggregation to any level above the block.

Evaluation of the operations and assumption of the estimators are given in Schenker (1988) and Hogan and Wolter (1988). Together with this paper, they demonstrate a thorough evaluation of the census counts and the undercount estimates of the test census.

ACKNOWLEDGEMENTS

I would like to thank the following people who contributed to this survey: Dan Childers, Howard Hogan, Cary Isaki, Matthew Jaro, Jan Jaworski, Charisse Jeffries, Robert O'Brien, Arona Pistiner, Nathaniel Schenker, Maria Urrutia, Debbie Wagner, and Kirk Wolter. I also would like to thank the referee for making comments that improved the presentation of this paper.

REFERENCES

- ANOLIK, IRWIN (1988). The 1986 rural post-enumeration survey in East Central Mississippi. Statistical Research Division Report Series. CENSUS/SRD/RR-87/09.
- CAUSEY, B.D., COX, L.H., and ERNST, L.R. (1985). Applications of transportation theory to statistical problems. *Journal of the American Statistical Association*, 80, 903-909.
- CITRO, CONSTANCE F., and COHEN, MICHAEL L. (1985). *The Bicentennial Census — New Directions for Methodology in 1990*. Washington: National Academy Press.
- ERICKSEN, EUGENE P., and KADANE, JOSEPH B. (1985). Estimating the population in a census year — 1980 and beyond. *Journal of the American Statistical Association*, 80, 98-109.
- FAY, R.E., PASSEL, J.S., and ROBINSON, J.G. (1988). The coverage of population in the 1980 census. 1980 Census of Population and Housing Evaluation and Research Report PHC80-E4, Washington: U.S. Government Printing Office.
- FREEDMAN, D. and NAVIDI, W. (1986). Models for adjusting the census. *Statistical Science*, 1, 3-11.
- HOGAN, HOWARD R. (1984). Research plan on adjustment. *Proceedings of the Social Statistics Section American Statistical Association*, 452-457.
- HOGAN, HOWARD R., and WOLTER, KIRK M. (1988). Measuring accuracy in a post-enumeration survey. *Survey Methodology*, 14.
- ISAKI, CARY T., SCHULTZ, LINDA K., SMITH, PHILIP J., and DIFFENDAL, GREGG J. (1987). Small area estimation research for census undercount-progress report. In *Small Area Statistics: An International Symposium*, 219-238. New York: John Wiley and Sons.
- JARO, MATTHEW, and CHILDERS, DANNY (1986). Matching the 1985 census of Tampa. Paper presented at the Annual Meeting of the American Statistical Association. Chicago.

- MARKS, ELI S. (1978). The use of dual system estimation in census evaluation. In *Developments in Dual System Estimation*, (ed. Karol Krotki), Edmonton: University of Alberta Press.
- MULRY, M.H., HOGAN, H.R., WALKER, J.R., CHAPMAN, D.R., EVAUL, J., and MOORE, R.H. (1981). Research proposal for a study of methods for 1990 decennial census coverage evaluation. Technical Report, U.S., Bureau of the Census, Washington, D.C.
- SCHENKER, NATHANIEL (1988). Handling missing data in coverage estimation, with application to the 1986 test of adjustment related operations. *Survey Methodology*, 14.
- U.S. Bureau of the Census (1979). Evaluation. In *POPSTAN: Program Considerations*, Part A, Chapter A-13.
- WOLFGANG, GLENN (1987). The pre-enumeration survey of the 1986 census of Central Los Angeles County. Paper presented at the Annual Meeting of the American Statistical Association, San Francisco.
- WOLTER, KIRK M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, 81, 338-346.
- ZASLAVASKY, ALAN M. (1988). Representing local area undercount by reweighting of households. Proceedings of the Bureau of the Census Fourth Annual Research Conference.

Handling Missing Data in Coverage Estimation, with Application to the 1986 Test of Adjustment Related Operations

NATHANIEL SCHENKER¹

ABSTRACT

This paper discusses methods used to handle missing data in post-enumeration surveys for estimating census coverage error, as illustrated for the 1986 Test of Adjustment Related Operations (Diffendal 1988). The methods include imputation schemes based on hot-deck and logistic regression models as well as weighting adjustments. The sensitivity of undercount estimates from the 1986 test to variations in the imputation models is also explored.

KEY WORDS: Imputation; Nonresponse; Post-enumeration survey; Weighting adjustments; Undercount.

1. INTRODUCTION

Missing data can be a major source of uncertainty in the estimation of coverage error for the decennial censuses in the United States (Freedman and Navidi 1986; Fay, Passel, and Robinson 1988, Chapter 6). For both the 1960 and 1980 Decennial Censuses, several estimates of coverage error were computed under different treatments of the missing data.

The Bureau of the Census has conducted many tests of methods for coverage error estimation to prepare to handle missing data and other problems for the 1990 Decennial Census. One such test was the 1986 Test of Adjustment Related Operations (TARO) (Diffendal 1988), which used the 1986 Census of Central Los Angeles County. Changes in field methodology and design for TARO reduced the levels of certain types of missing data from the levels for 1980 (Hogan and Wolter 1988). Nevertheless, some missing-data problems remained.

This paper describes the missing-data problems in TARO and how they were handled in the estimation process. Section 2 gives a brief description of how coverage error was estimated in TARO. Sections 3-6 discuss the types of missing data that occurred, the extent to which they occurred, and the methods used to handle them. These methods include a weighting adjustment for unit nonresponse (noninterviews), hot-deck imputation for missing demographic and housing characteristics, and imputation using logistic regression models for certain binary items related to enumeration in the census. Section 7 presents coverage error estimates under alternative imputation models and alternative treatments of certain problem cases. The lowest and highest estimated undercount rates obtained using these alternatives are 8.50% and 10.16% for Hispanics, 5.86% and 7.81% for Asian non-Hispanics, and 5.81% and 6.59% for Others. The estimates from TARO for the three race categories were 9.85%, 7.32%, and 6.21%, respectively. A concluding discussion is given in Section 8.

2. ESTIMATING CENSUS COVERAGE ERROR

Diffendal (1988) discusses in detail how census coverage error was estimated in TARO. This section describes briefly those aspects necessary for understanding the rest of this paper.

¹ Nathaniel Schenker, Undercount Research Staff, Statistical Research Division, Bureau of the Census, Washington, DC 20233, USA. This paper reports research undertaken by a member of the Census Bureau's staff. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau.

Coverage error was estimated using data from a post-enumeration survey (PES) of people in the census site. First a sample of blocks in the site was drawn. Then each housing unit in the sample blocks was surveyed to determine its occupants on Census Day, its occupants at the time of the PES and where they lived on Census Day, and the characteristics of the occupants.

Two samples were used to estimate census coverage error. The P (population) sample was composed of the people who lived in the PES sample blocks at the time of the PES. An attempt was made to match each P-sample person to a person enumerated in the census to determine whether the P-sample person had been enumerated; the match rate within each domain of study was used essentially to estimate the capture rate of the census for that domain. The E (enumeration) sample was composed of the people who were enumerated in the census as living in the PES sample blocks; this sample was used to estimate the number of erroneous enumerations (e.g., fictitious enumerations and duplicates) and unmatchable persons (e.g., persons for whom no names were reported) in the census within each domain. An attempt was made to match each E-sample person to a person in the PES. Each E-sample match was considered a correct enumeration since the PES indicated that the person should have been enumerated. Each E-sample nonmatch was followed up to determine whether it was an erroneous enumeration or a correct enumeration that was missed in the PES (which is not itself assumed to have perfect coverage).

If a PES of the entire United States were conducted, individuals in the P-sample who moved out of Central Los Angeles County between Census Day and the PES would be interviewed in the PES. An attempt to match these individuals to census enumerations in Central Los Angeles County would be made, and the resulting data would be used in the estimation of coverage error for Central Los Angeles County. Similarly, individuals in the P-sample who moved into Central Los Angeles County between Census Day and the PES would contribute to coverage error estimates outside of Central Los Angeles County. However, because the census and PES for TARO were conducted only in Central Los Angeles County and not in the entire United States, outmovers from the test site were not interviewed in the PES and inmovers did not apply to the test. Thus data for inmovers and outmovers were not used in the estimation. (Note that data for movers *within* test site were used, however). This issue is discussed further in Section 7.2.

The "dual-system" estimator of the population size (see Marks, Seltzer, and Krotki 1974, Krotki 1978, and Wolter 1986 for discussion and references) is written

$$DSE = N_p (CEN - SUB - EE) / M, \quad (1)$$

where N_p is the weighted number of people in the P-sample, CEN is the unadjusted census count, SUB is the number of whole-person substitutions (for unit nonresponse) in the census, EE is a weighted estimate of the number of erroneous enumerations and unmatchable persons in the census, and M is the weighted number of matches between the P-sample and census; census data provide CEN and SUB, whereas P- and E-sample data provide N_p , EE, and M. The dual-system estimator can be thought of as inflating the estimated number of correct and matchable census enumerations (CEN-SUB-EE) by the inverse of the estimated census capture rate (M/N_p).

The theory of dual-system estimation assumes that for both the census and the PES, the probability of capture is constant across all people in the domain to which the estimator is applied (Wolter 1986). Thus no one group of people in the domain should be more or less likely to be enumerated in the census or PES than any other group. To make this assumption more realistic in TARO, separate dual-system estimates were computed within poststrata based on

person and housing characteristics. The poststrata are described in Diffendal (1988). One example is the Hispanic male renters of ages 30 to 44 living in primarily Hispanic blocks.

To summarize, the P- and E-sample data needed for coverage error estimation were the match status (match vs. nonmatch) for each P-sample person, the enumeration status (correct vs. erroneous) for each E-sample person, and person and housing characteristics for each person in both samples.

3. P-SAMPLE HOUSEHOLD NONINTERVIEWS

Occasionally, a PES interviewer was unable to obtain an interview for an occupied housing unit; this occurred, for example, when the occupants refused to respond. Of the 5,935 housing units that were judged to be nonvacant, 32 (0.5%) were classified as having household noninterviews. The occurrence of household noninterviews resulted in missing data on the number of people in each household, person and housing characteristics, and match statuses.

The block-sample design of the PES afforded a simple way to handle P-sample household noninterviews. Within each sample block, the sampling weights of the noninterview households were redistributed across the interviewed households. The noninterview weighting adjustment basically assumes that the distributions of people, characteristics, and match statuses for households not interviewed within a block are the same as for households interviewed. This assumption was used because households tend to be more similar within blocks than across blocks, although noninterview households still probably differ somewhat from interviewed households, especially with respect to household size (see, e.g., Palmer 1967).

It is possible that the data obtained for a household by proxy interview (which in TARO referred to a completed interview with someone outside the household) are of sufficiently low quality that such a household should be classified as a noninterview household. The quality of data from the 189 proxy interviews in TARO is discussed in Section 4, and some coverage error estimates with proxy interviews treated as noninterviews are presented in Section 7.

4. MISSING CHARACTERISTICS IN THE P- AND E-SAMPLES

Even when an interview was obtained for a P-sample household, the data on person and housing characteristics were sometimes incomplete. Incomplete data on characteristics also occurred in the census and therefore in the E-sample.

The variables used in poststratification for TARO (Diffendal 1988) included the housing variable Tenure (1 = owned, 2 = rented or occupied without payment) and the person variables Sex (1 = male, 2 = female), Age (1 = 0-14, 2 = 15-29, 3 = 30-44, 4 = 45-64, 5 = 65+), and Race (1 = Hispanic, 2 = Asian non-Hispanic, 3 = Other). In addition, the housing variable Structure (1 = single-unit, 2 = multiunit) was used in handling missing P-sample match statuses and missing E-sample enumeration statuses (see Sections 5 and 6).

Table 1 displays the missing-characteristic counts for the entire P- and E-samples and for cases coming from P-sample proxy interviews. For the P- and E-samples, the highest missing-data rate was 7.0% for E-sample Race, with all other rates being 3.5% or lower. The missing-data rates for P-sample proxy cases were all several times higher than those for the entire P-sample, although only Tenure (20.2%) had a rate higher than 10%.

Missing characteristics for each of the samples (P and E) were imputed by a hot-deck method involving two passes through the data after the data had been sorted geographically. On the first pass, missing values of Tenure, Structure, and Race were imputed using the most

Table 1
Missing-Characteristic Counts (% in Parentheses)
for the Entire P- and E-Samples and for P-Sample Proxy Interviews

Variable	P-Sample (19,552 persons)	E-Sample (20,976 persons)	P-Sample Proxy (430 persons)
Tenure	690 (3.5)	154 (0.7)	87 (20.2)
Structure	459 (2.3)	343 (1.6)	38 (8.8)
Sex	418 (2.1)	82 (0.4)	18 (4.2)
Age	137 (0.7)	432 (2.1)	18 (4.2)
Race	155 (0.8)	1463 (7.0)	17 (4.0)

NOTE: The 19,552 persons in the P-sample include the 430 proxy cases.

recent observed data, because of the presumed strong relation between these variables and geography. In addition, distributions of Sex and Age were tabulated for categories of type of household (single-person vs. multiperson), marital status, relationship to head of household, and sex and age of head of household, using all observed data. On the second pass, missing values of Sex and Age were imputed at random from the distributions tabulated during the first pass. Further details on the imputation of characteristics in TARO can be found in Schenker (1987).

In summary, the block sample design of the PES was helpful not only in developing a noninterview weighting scheme (Section 3), but also in the imputation of characteristics that tend to be clustered by block, that is, Tenure, Structure, and Race.

5. MISSING MATCH STATUSES IN THE P-SAMPLE

Of the 19,552 P-sample cases resulting from completed interviews, 161 (0.8%) were missing match statuses for dual-system estimation. All but three of these unresolved cases fell into two broad categories: 105 cases for which matching was not attempted due to incomplete names and/or insufficient characteristics; and 53 movers between Census Day and the PES for whom there were problems specifying a Census Day address or finding the census questionnaire for the Census Day address.

A traditional approach to handling a missing binary item such as match status is to impute one of the two possible outcomes for the missing item. For example, in the estimation of undercount for the 1980 Decennial Census, the match status for each unresolved P-sample case was imputed from a resolved case with similar characteristics (Fay, Passel, and Robinson 1988, Chapter 6). A different approach was taken in TARO, however. After all missing characteristics were imputed using the methods described in Section 4, a match probability was imputed for each unknown match status; the probability was estimated using an explicit model (to be described later in this section). The contribution of the unresolved cases to the M term of the dual-system estimate (1) was the weighted sum of the imputed probabilities.

Probabilities rather than binary outcomes were imputed for two reasons. First, imputing random binary outcomes is less efficient than imputing estimated probabilities, yielding estimates with higher variances (see Rubin 1987, p. 15). Second, because imputed probabilities represent uncertainty about the missing match statuses, it should be possible to use the probabilities to obtain a variance due to imputation. Note, however, that since the dual-system estimator (1) is nonlinear in M, imputing a probability (or mean) for each missing binary

outcome introduces some bias into the estimation (see Rubin 1987, p. 14). Current research is investigating the use of imputed probabilities for missing binary data.

The following logistic regression approach was used to impute match probabilities. Let X denote a vector of predictors, $Y = \text{match or nonmatch}$, and $p = \Pr(Y = \text{match} | X)$. The parameter vector β of the logistic regression model

$$\text{logit}(p) = \log[p / (1 - p)] = X' \beta$$

was estimated from the data for the resolved cases using the Bayesian techniques for categorical logistic regressions described in Rubin and Schenker (1987); these techniques involve adding fractional observations to each cell in the logistic regression and then fitting the model by standard maximum-likelihood methods. Then for unresolved case j , with $X = x_j$, the imputed match probability was

$$\hat{p}_j = \text{logit}^{-1}(x_j' \hat{\beta}) = \exp(x_j' \hat{\beta}) / [1 + \exp(x_j' \hat{\beta})],$$

where $\hat{\beta}$ denotes the estimate of β . The background variables used to define X were Tenure, Structure, Sex, Age, and Race, as well as variables indicating regular interview versus proxy interview and mover versus nonmover between Census Day and the PES.

Table A1 (in the Appendix) gives the logistic regression coefficient estimates. The large coefficients associated with interview and mover status indicate that proxy and mover cases have much lower imputed match probabilities than others. It may be that these lower match probabilities are due in part to difficulties in matching proxy and mover cases rather than just lower census capture rates for these cases. If this is true, alternative treatments of the data may be in order; such alternatives are considered in Section 7.

Of the 19,391 resolved P-sample cases, 17,018 (87.8%) were matches. The (unweighted) sum of the 161 imputed match probabilities was 124.66; thus the imputed match rate was 77.4%. Although a stratified sample of blocks was used in TARO, the estimation of the logistic regression parameters assumed a simple random sample of people. To examine the possible biases due to not accounting for the stratification, the logistic regression was fitted again (after TARO was completed) with indicator variables for the six sampling strata (Diffendal 1988) included in X . The result of this refinement is a sum of imputed match probabilities equal to 124.50 (77.3%). The minor effect of this change on estimates of census coverage error is demonstrated in Section 7. Implications of possible design effects due to clustering are discussed in Section 8.

6. MISSING ENUMERATION STATUSES IN THE E-SAMPLE

Of the 20,976 cases in the E-sample, 3,714 were followed up or should have been followed up. After followup, 979 cases (4.7% of total, 26.4% of followup) had missing enumeration statuses. All but nine of these unresolved cases fell into four broad categories: 498 cases that should have been followed up but were not; 257 cases in which the respondent to the followup interview did not know the person in question; 137 cases for which the interview yielded insufficient information to determine an enumeration status; and 78 cases for which there were followup noninterviews.

Missing enumeration statuses in the E-sample were handled by imputing a probability of erroneous enumeration for each unresolved case. The contribution of the unresolved cases to the EE term of the dual-system estimate (1) was the weighted sum of the imputed probabilities. The imputation procedure was analogous to that used for P-sample match statuses with one

major change: Since missing enumeration statuses resulted solely from followup, only the resolved cases from followup were used in estimating the logistic regression. The background variables used to define X for the logistic regression were Tenure, Structure, Sex, Age, and Race, along with variables indicating whether the census questionnaire for the person's household was returned by mail and whether the entire household or only part of the household was not matched before followup. Table A2 (in the Appendix) gives the logistic regression coefficient estimates.

Of the 17,262 non-followup cases, 278 (1.6%) were classified as erroneous enumerations or unmatchable. There were 2,735 resolved followup cases, of which 82 (3.0%) were classified as erroneous enumerations. The (unweighted) sum of the 979 imputed probabilities was 21.93 (2.2%). When indicator variables for the sampling strata are included in X , the sum changes to 23.58 (2.4%). As with the P-sample, this change has a very minor effect on estimates of coverage error; see Section 7.

7. ESTIMATES OF COVERAGE ERROR UNDER ALTERNATIVE TREATMENTS OF MISSING DATA AND OTHER PROBLEM CASES

This section examines the effects of alternative treatments of missing data and other problem cases on estimates of coverage error for the three categories of race defined by the variable Race (Hispanic, Asian non-Hispanic, and Other). For a given treatment and race category, let \hat{N} be the sum of the dual-system estimates over all poststrata corresponding to the race category and let N_c be the sum of the unadjusted census counts over the poststrata. The estimated undercount rate is then $100(1 - N_c / \hat{N})\%$.

Consider first the alternative of including indicators of the sampling strata as predictors in the P- and E-sample logistic regressions for imputing match and erroneous enumeration probabilities, as discussed in Sections 5 and 6. The estimated undercount rates from TARO, which were obtained without using these predictors, are 9.85% for Hispanics, 7.32% for Asian non-Hispanics, and 6.24% for Others. When indicators of the sampling strata are used, the estimates change to 9.82% for Hispanics, 7.31% for Asian non-Hispanics, and 6.21% for Others. The largest difference due to including the sampling stratum indicators is only 0.03%. For all the alternative treatments to be considered, however, this refinement is used because it is in principle more correct; for instance, it should yield more accurate standard errors.

7.1 Treatments that Lower the Estimated Undercount

The match rate for the 375 resolved P-sample proxy cases was 78.9% as opposed to the overall P-sample rate of 87.8%. While it may be true that proxy cases were actually captured in the census less frequently than others, it is possible that part of the difference in the match rates is due to missing and/or incorrect proxy data (see Section 4). A conservative treatment would be to classify the 189 proxy interviews as household noninterviews and apply the weighting adjustment described in Section 3; this would essentially assign proxy cases the same match rate as nonproxy cases. (Note that when all proxy interviews are classified as noninterviews, an indicator of proxy/nonproxy status is no longer included in the logistic regression model for imputing match probabilities).

The match rate for the 277 resolved P-sample movers (between Census Day and the PES) was 66.1%. It is generally believed that movers are captured in the census at a lower rate than nonmovers, but it may be that the low match rate for movers is partly due to difficulties inherent in matching movers, such as problems in obtaining a correct Census Day address. A conservative

Table 2
Estimated Undercount Rates (in %) by Race Under Alternative Treatments
of P-sample Proxy Interviews, P-sample Movers, and E-sample W1's

Treatment (1 = alternative, 0 = TARO)			Hispanic	Asian non-Hispanic	Other
Proxy	Mover	W1			
0	0	0	9.82	7.31	6.21
0	0	1	9.30	6.76	5.83
0	1	0	9.33	7.24	6.19
0	1	1	8.80	6.69	5.81
1	0	0	9.55	6.52	6.24
1	0	1	9.03	5.96	5.86
1	1	0	9.04	6.45	6.22
1	1	1	8.51	5.90	5.84

NOTE: Indicators of the sampling strata were used as predictors in the logistic regressions for imputing match and erroneous enumeration probabilities.

treatment would be to classify all cases for movers as unresolved and then impute match probabilities for unresolved cases using a logistic regression model that does not include mover/nonmover status as a predictor. This would essentially assign movers the same match rate as nonmovers.

Of the 979 unresolved E-sample cases, 257 had the followup interview code W1, meaning that the respondent did not know the person in question. A code of W1 could have indicated that the person in question was fictitious. Therefore, after TARO, all W1's were reviewed by experienced matching personnel. Any case that showed evidence (such as a note from the interviewer) of possibly being fictitious was marked; there were 118 such cases. An alternative treatment to that used in TARO would be to classify the 118 cases as resolved erroneous enumerations before imputation. This would raise both the observed and imputed rates of erroneous enumeration.

Table 2 displays the undercount estimates by race category for the 2x2x2 factorial design with the factors being whether or not alternative treatments are used for proxy interviews, movers, and W1's. The ranges between the lowest and highest estimated undercount rates are 1.31% for Hispanics, 1.41% for Asian non-Hispanics, and 0.43% for Others.

Note that for each race category, there is not much interaction between the treatments of proxy interviews, movers, and W1's. In fact, the following simple additive model can be used to predict the entries in Table 2 for each race category:

$$\hat{Y} = \hat{\alpha}_0 + I_p \hat{\alpha}_p + I_m \hat{\alpha}_m + I_w \hat{\alpha}_w, \tag{2}$$

where \hat{Y} is the predicted estimate of the undercount rate, I_p , I_m , and I_w are the treatment indicators (1 = alternative, 0 = TARO) for proxy interviews, movers, and W1's, respectively, and $\hat{\alpha}_0$, $\hat{\alpha}_p$, $\hat{\alpha}_m$, and $\hat{\alpha}_w$, are parameter estimates given in Table 3. The parameter α_0 is the estimated undercount rate when no alternative treatments are used; α_p , α_m , and α_w are the effects of using alternative treatments for proxy interviews, movers, and W1's, respectively. The largest residual when equation (2) is used to predict the entries in Table 2 is 0.02%.

Table 3
Parameter Estimates for the Additive Model (2) for Predicting
the Estimated Undercount Rates in Table 2

	Hispanic	Asian non-Hispanic	Other
$\hat{\alpha}_o$	9.82	7.31	6.21
$\hat{\alpha}_p$	-0.28	-0.7925	0.03
$\hat{\alpha}_m$	-0.505	-0.0675	-0.02
$\hat{\alpha}_w$	-0.525	-0.5525	-0.38

7.2 A Procedure that Raises the Estimated Undercount

Because TARO was confined to one small area in the United States, no PES data could be obtained for people who moved out of the test site between Census Day and the PES. The omission of these outmovers from estimation was equivalent to assuming that they had the same capture rate in the census as the included cases. This was a conversative assumption, since movers are generally believed to have a lower capture rate than nonmovers.

There were 409 people who moved into the test site between Census Day and the PES. These in-movers were not included in the estimation because their Census Day addresses were outside the test site and thus their data applies to other areas. Moreover, there were no census cases to which to match the in-movers since they were outside the test site on Census Day.

A procedure that might indicate the effect of including outmovers in the estimation would be to include the 409 in-movers as substitutes and impute match probabilities for them (since their match statuses are unknown). The treatments yielding the highest and lowest estimates in Table 2 have been applied to the TARO data with in-movers included; the results are displayed in Table 4. Note that the lower estimated undercount rates in Table 4 (obtained using the alternatives to the TARO treatments for proxy interviews, movers, and W1's) are all within 0.04% of the corresponding estimates in Table 2. This result is expected, since the addition of cases having an imputed match rate that is approximately the same as the overall match rate should not affect the estimates much. The higher estimates in Table 4 are larger than the corresponding estimates in Table 2 by 0.34% for Hispanics, 0.50% for Asian non-Hispanics, and 0.38% for Others.

Table 4
Estimated Undercount Rates (in %) by Race When In-movers are
Included in the Data with Imputed Match Probabilities

Treatment (1 = alternative, 0 = TARO)			Hispanic	Asian non-Hispanic	Other
Proxy	Mover	W1			
0	0	0	10.16	7.81	6.59
1	1	1	8.50	5.86	5.81

NOTE: Indicators of the sampling strata were used as predictors in the logistic regressions for imputing match and erroneous enumeration probabilities.

8. SUMMARY AND DISCUSSION

A combination of weighting and (random and nonrandom) imputation methods was used to handle missing data in TARO. P-sample household noninterviews were handled by a block-level weighting adjustment. A hot-deck imputation method was used for missing characteristics in both samples. Missing P-sample match statuses and E-sample enumeration statuses were handled using imputed probabilities estimated by logistic regression methods.

As mentioned in Sections 5 and 6, the use of imputed probabilities for missing P-sample match statuses and E-sample enumeration statuses should facilitate the assessment of variability due to imputing these statuses. To assess this variability completely, it is necessary to measure variability due to estimating the logistic regression parameters as well as the variability due to imputation given β (Rubin and Schenker 1986). Thus an estimated variance-covariance matrix for $\hat{\beta}$ is needed. Since a cluster sample was used in TARO, the logistic regression estimation procedures (Section 5), which assume a simple random sample, do not provide an accurate estimate of the variance-covariance matrix. This was not a major concern in TARO, because the measurement of imputation variance was not a primary goal. Moreover, for the nonresponse rates achieved in TARO, the variability due to uncertainty in estimating β is likely to be minor relative to the uncertainty due to imputation given β (Rubin and Schenker 1986).

Although it is possible in principle to assess the variability due to imputing match and enumeration statuses using the TARO procedures, variability due to imputing missing characteristics (Section 4) cannot be quantified. One way to make the quantification of such variability possible would be to multiply impute characteristics in the P- and E-samples (Rubin 1987). Several dual-system estimates would then need to be calculated, however — one for each set of imputations.

The models underlying the weighting and imputation methods used in TARO assume that given the observed data, the chance of a variable being missing does not depend on its value. Another issue regarding imputation is how best to impute characteristics and match statuses (or enumeration statuses) simultaneously. The TARO procedure of first imputing characteristics and then imputing statuses conditional on the imputed characteristics assumes that statuses are not useful predictors for imputing characteristics. Models that relax the TARO assumptions may be more appropriate. Rubin, Schafer, and Schenker (1988) discuss this further.

Missing data are only one source of error in estimating coverage. Other sources, such as matching error and violations of the assumption of constant capture probabilities (Section 2), are discussed in Hogan and Wolter (1988). After assessing all of these sources of error for TARO, Hogan and Wolter conclude that the TARO coverage measurement is more accurate than the original enumeration.

ACKNOWLEDGEMENTS

I would like to thank Robert O'Brien for computing the dual-system estimates used in Section 7. I am also grateful to the referees for helpful comments and to the members of the workshop on the undercount of the Harvard University Statistics Department for stimulating discussions on handling missing data in undercount estimation.

APPENDIX

LOGISTIC REGRESSION RESULTS

Table A1
Results for P-Sample Logistic Regression

Predictor	Codes	Estimated Coefficient
Intercept		1.47
Interview Status	1 if regular, - 1 if proxy	.36
Mover Status	1 if nonmover, - 1 if mover	.60
Tenure	1 if owner, - 1 otherwise	.46
Structure	1 if single-unit, - 1 if multiunit	-.16
Sex	1 if male, - 1 if female	-.09
Age 1	1 if 0-14, - 1 if 65 +, 0 otherwise	-.06
Age 2	1 if 15-29, - 1 if 65 +, 0 otherwise	-.46
Age 3	1 if 30-44, - 1 if 65 +, 0 otherwise	-.02
Age 4	1 if 45-59, - 1 if 65 +, 0 otherwise	.13
Race 1	1 if Hispanic, - 1 if Other, 0 if Asian non-Hispanic	-.14
Race 2	1 if Asian non-Hispanic, - 1 if Other, 0 if Hispanic	.11

Table A2
Results for E-Sample Logistic Regression

Predictor	Codes	Estimated Coefficient
Intercept		- 3.45
Questionnaire Status	1 if mail-return, - 1 otherwise	.01
Pre-followup Status	1 if partial-household match, - 1 if whole-household nonmatch	-.20
Tenure	1 if owner, - 1 otherwise	.36
Structure	1 if single-unit, - 1 if multiunit	.17
Sex	1 if male, - 1 if female	.08
Age 1	1 if 0-14, - 1 if 65 +, 0 otherwise	-.30
Age 2	1 if 15-29, - 1 if 65 +, 0 otherwise	-.04
Age 3	1 if 30-44, - 1 if 65 +, 0 otherwise	-.34
Age 4	1 if 45-59, - 1 if 65 +, 0 otherwise	.10
Race 1	1 if Hispanic, - 1 if Other, 0 if Asian non-Hispanic	-.02
Race 2	1 if Asian non-Hispanic, - 1 if Other, 0 if Hispanic	-.38

REFERENCES

- DIFFENDAL, G. (1988). The 1986 Test of Adjustment Related Operations in Central Los Angeles County, in this issue.
- FAY, R.E., PASSEL, J.S., and ROBINSON, J.G. (1988). *The Coverage of Population in the 1980 Census*, 1980 Census of Population and Housing Evaluation and Research Report PHC80-E4, Washington: U.S. Government Printing Office.
- FREEDMAN, D.A., and NAVIDI, W.C. (1986). Regression Models for Adjusting the 1980 Census, *Statistical Science*, 1, 3-39.
- HOGAN, H.R., and WOLTER, K.M. (1988). Measuring Accuracy in a Post-Enumeration Survey, in this issue.
- KROTKI, K. (1978). *Developments in Dual System Estimation of Population Size and Growth*, Edmonton: The University of Alberta Press.
- MARKS, E.S., SELTZER, W., and KROTKI, K.J. (1974). *Population Growth Estimation*, New York: The Population Council.
- PALMER, S. (1967). On the Character and Influence of Nonresponse in the Current Population Survey, *Proceedings of the Social Statistics Section*, American Statistical Association, 73-80.
- RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.
- RUBIN, D.B., SCHAFER, J.L., and SCHENKER, N. (1988). Imputation Strategies for Estimating the Undercount, *Proceedings of the Fourth Annual Research Conference*, U.S. Bureau of the Census.
- RUBIN, D.B., and SCHENKER, N. (1986). Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse, *Journal of the American Statistical Association*, 81, 366-374.
- RUBIN, D.B., and SCHENKER, N. (1987). Logit-Based Interval Estimation for Binomial Data Using the Jeffreys Prior, *Sociological Methodology*, 17, 131- 144.
- SCHENKER, N. (1987). Handling Missing Data in the 1986 Test of Adjustment Related Operations, *Proceedings of the Section on Survey Research Methods*, American Statistical Association.
- WOLTER, K.M. (1986). Some Coverage Error Models for Census Data, *Journal of the American Statistical Association*, 81, 338-346.

Measuring Accuracy in a Post-Enumeration Survey

HOWARD HOGAN and KIRK WOLTER¹

ABSTRACT

The U.S. Bureau of the Census will use a post-enumeration survey to measure the coverage of the 1990 Decennial Census. The Census Bureau has developed and tested new procedures aimed at increasing the accuracy of the survey. This paper describes the new methods. It discusses the categories of error that occur in a post-enumeration survey and means of evaluation to determine that the results are accurate. The new methods and the evaluation of the methods are discussed in the context of a recent test post-enumeration survey.

KEY WORDS: Census; Undercount; Overcount; Coverage Evaluation.

1. INTRODUCTION

In this article we discuss recent research at the U.S. Bureau of the Census to improve the accuracy of a post-enumeration survey and to measure that accuracy. Much of this research was originally directed toward the goal of developing a sound body of statistical theory, methods, and operations for correcting U.S. census figures for coverage errors. The results presented in this paper show that we are now able to produce PES estimates of total population that are closer to the true population than are original census estimates.

In light of a policy decision made by the U.S. Department of Commerce not to correct the 1990 enumeration for coverage error, the PES methods we discuss will be used to provide a careful evaluation of the coverage of the 1990 Census. See U.S. Department of Commerce (1987). This evaluation will be used to inform users of the limitations of the census, to inform planning for future censuses, or to improve the Census Bureau's estimates of the U.S. population for years subsequent to the census year.

The PES method uses two samples to measure net coverage error. A sample of people who should have been counted in the original census enumeration is interviewed after the census and is used to measure census omissions. We call this the population or "P" sample. One also needs a sample of census enumerations to measure duplicates and other errors included in the census count. We call this the enumeration or "E" sample. The samples form an estimate of total population using the dual system-estimator (DSE). See Diffendal (1988) for a full discussion of the samples and the dual-system model. Unless otherwise stated, we will use Diffendal's notation throughout this article.

The Census Bureau conducted a PES in conjunction with the 1980 Census. The P sample consisted of persons in households enumerated in the April and August Current Population Survey (CPS) samples. For a description of the CPS, see U.S. Bureau of the Census (1978). The E sample was a separate and independent sample of persons in housing units enumerated in the census. In addition, the Census Bureau produced an alternative set of undercount estimates based upon an aggregate analysis of birth and death registration data, administrative

¹ U.S. Bureau of the Census, Washington D.C. 20233. Howard Hogan is Chief of the Undercount Research Staff. Kirk Wolter is Chief of the Statistical Research Division. This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributed to the authors and do not necessarily reflect those of the Census Bureau.

records, and previous censuses. This program, called demographic analysis, will be referred to occasionally in this article. The Census Bureau did not correct the 1980 enumeration for undercount errors because we considered the PES estimates to be flawed by missing and inaccurate data. In addition, the demographic analysis results were flawed by, among other things, a lack of data on the number of undocumented immigrants and the lack of an acceptable method to carry the estimates down to the state and local level. See Fay *et al.* (1988).

In very recent years, we have developed a new PES design and new methodology that minimizes the problems experienced in 1980, while not creating major new ones. The new PES design is based on a common area sample of census blocks for both the P and E samples. The P sample consists of all people living in the sample blocks at the time of PES interviewing. Interviewers visit each housing unit and determine where the residents were living at the time of the census.

Using newly developed computer matching methods and software (Jaro 1988), we attempt to match all P-sample people to corresponding census enumerations. Clerks review the computer's work and make a final determination as to the enumeration status (either enumerated or missed in the original enumeration) of each P-sample person. For people who moved between the census and the PES, we assign the census-day address to the proper block and search for a match there. For a few cases, matching is indeterminate at this point, and a further interview or followup is necessary either to gather additional information or to resolve conflicts in existing information. After the followup, clerks assign an enumeration status to the P-sample people for whom the followup interview is complete. For a very few residual cases, matching may be still unresolved, and we impute to each an enumeration status, using appropriate statistical techniques for missing data (Schenker 1988).

For each E-sample person, a determination is made as to the person's enumeration status (either correctly enumerated or erroneously enumerated) in the original census. Section 6 gives a description of what constitutes an erroneous enumeration (EE), and all non-erroneous enumerations are considered correct enumerations (CE). In many cases, the census enumerates the same people that are interviewed as part of the P sample. Thus, the two samples overlap to a great extent. Most E-sample people who are also in the P sample (as determined by the computer and clerical matching system) are automatically declared CE. However, the overlap is not complete. The P sample will miss some people that are included in the E sample and vice versa. The census will enumerate others in the block by mistake. Interviewers will invent some enumerations. For all E-sample people who are not matched to a P-sample person, it is necessary to conduct a followup interview. This followup gathers enough information to allow a determination of whether the E-sample people were counted correctly in the original enumeration.

We tested the new PES design in 1986 in connection with a test census in Los Angeles. The test was called the Test of Adjustment Related Operations (TARO) and consisted of 190 blocks, containing almost six thousand housing units and 20,000 people. The estimated net undercount for the Los Angeles test was about 9 percent. For details on TARO methods and results, see Diffendal (1988) and Schenker (1988).

We also tested the new PES design in a rural area of Mississippi during 1986. There we used a sample of 271 blocks with about 3250 housing units and eight thousand people. The estimated undercount in this test was 5.5 percent. For details of results and methodology, see Anolik (1988). Although, the Mississippi test data have not been as completely analyzed as the TARO data, we will refer occasionally to the results in this article.

An important question is whether the new PES can produce more accurate estimates of population than can the original census enumeration. In theory, the PES estimates should be considered the more accurate, but in practice, nonsampling errors can and do arise in the

Table 1
TARO Errors and Estimates of the Mean Effect on the
Estimated Undercount of Correcting the Error

Sources of Error	Mean Effect on Estimated Undercount
Matching error	- 1.0%
Reporting census-day address	- 1.0%
Fabrication in the PES interview	- 1.0%
Missing data	0.0%
Error in measuring the erroneous enumerations	- 0.5%
Balancing gross overcounts and undercounts	0.0%
Correlation bias	+ 2.3%
Random error	0.0%

conduct and analysis of both the PES and the census enumeration. Careful study is needed to assess their relative accuracies. In this article, we present our assessment of the error structure of the 1986 TARO.

Eight potential sources of error affect coverage measurements produced by the PES: sampling error plus seven sources of nonsampling error. The sources and our summary assessment of their impact on TARO data are presented in Table 1. The second column gives the effects of the errors on the estimated undercount. For example, if we correct all "matching errors," the estimated undercount would be reduced by about one percentage point, from 9 percent to 8 percent. Some errors, such as "missing data" and "random error", might either raise or lower the undercount, and our best assessment is that these errors introduce no important bias into TARO data. The figures in this column represent assessments of individual error, without regard for the other sources of error.

By construction, the eight individual errors tend to be mutually exclusive and additive. Some overlaps or interactions are possible between the different sources, but we believe they are small and we ignore them here. Overall, we calculate the joint effect of the errors as

$$(-1.0 - 1.0 - 1.0 + 0.0 - 0.5 + 0.0 + 2.3 + 0.0) \text{ percent} = -1.2 \text{ percent.}$$

Thus, correcting for the joint effect of the errors would lower the estimated undercount from 9.0 percent to about 7.8 percent. The corrected figure, 7.8 percent, may be viewed approximately as the mean of a posterior error distribution for the TARO undercount. Development of a complete posterior error distribution is proceeding at the Census Bureau (see Mulry and Spencer 1988).

Because the original TARO estimate of 9 percent is much closer to the corrected figure of 7.8 percent than the corrected figure is to zero, we conclude that the original TARO data is closer to the truth than is the original census enumeration.

In the next 8 sections of the article, we treat the error components one by one. Each section discusses both the procedures and problems confronted in the 1980 PES, and the error-resistant improvements that were tested in TARO. We describe the evaluation of each error component and the evidence for our conclusions. The paper closes in Section 10 with a summary of our findings and some directions for future research.

2. MATCHING ERROR

Errors in classifying P-sample people as enumerated or not can occur for two general reasons:

- (a) the information reported by the respondent/interviewer is incorrect
- (b) correct information is reported, but not correctly used.

Category (a) consists of errors in the reporting of census-day address and fabrication in the PES interview, discussed in Sections 3 and 4, respectively. The present section discusses matching errors (category (b)) that occur even when the people are real and their census-day address is correctly reported. In other words, these are errors in matching due to processing mistakes.

In our new PES design, matching takes two forms: automated batch matching and computer-assisted clerical matching. The status of "not enumerated" is assigned to a P-sample person when sufficient information for matching has been gathered and no matching case can be found in the census. Errors occur when there actually was insufficient information for matching but matching was attempted nonetheless, and also when the correct census questionnaires were searched but the match was not established, even though the person was in fact counted in the original enumeration.

A P-sample person occasionally may be declared to match the wrong census person. This happens most often within families, where children's names and ages may be similar, and in "ethnic" neighborhoods where certain names are unusually common. Normally, false matches are less common than false nonmatches because the matches can be reviewed easily by a clerical matching staff. False matches create a bias in the dual system estimator only when the P-sample person was actually not enumerated.

A principal change in our PES design since 1980 that allows better control of matching error is the use of a common sample of blocks for both P and E samples. The block sample design permits a classification of all enumerated people (both P-and E-sample) into three categories:

- counted in P sample, counted in E sample
- counted in P sample, missing from E sample
- missing from P sample, counted in E sample.

This kind of organization or accounting, which was not possible with the 1980 design, imparts to the matching process a quality that resists matching error. For example, people with similar names in ethnic neighborhoods can be sorted out using all the information provided by a block sample. Address mix-ups in the census process are easier to handle with a block sample. The choice of census block as a sampling unit also reduces geographic coding error as compared to the 1980 PES, where the P sample was based on CPS clusters of four housing units and 1970 Census geography.

Matching is especially difficult for P-sample people who lived elsewhere on census day, i.e., movers. For movers, the census-day address reported in the P-sample interview must be assigned to the proper geographical area prior to matching. This assignment was problematic in the 1980 PES and the new design does not necessarily solve the problem. The Census Bureau will, however, be using a new, automated geographical system for the 1990 Census (see Marx and Saalfeld 1988), and we are hopeful that this innovation will permit rapid and accurate geographic assignment for mover addresses.

In the 1986 TARO, about 74 percent of the P-sample people were matched by the computer. Another 12 percent were declared "possible match" by the computer. A specially trained clerical staff reviewed all cases not designated as "match" by the computer, including all of the computer-designated "possible matches."

Table 2
Results of Rematch Study: Sample (Weighted)^a

Results of Original Matching	Results of Rematching			Total
	Enu- merated	Not Enu- merated	Un- resolved	
Enumerated	16,623	18	55	16,696
Not Enumerated	88	2,164	56	2,308
Unresolved	17	0	132	149
Total	16,728	2,182	243	19,153

^a Weighting is to P-sample totals.

The results of the 1986 PES in Mississippi show that the success of the computer matching system is not limited to urban areas with house numbers, street names and well-defined geography. In the Mississippi test, addresses commonly consisted of a rural route and box number. Blocks were irregularly shaped with invisible boundaries such as an intermittent stream or county line. Still, the computer was able to match 68 percent of the cases.

We have conducted two studies to evaluate the extent of matching error in TARO. In the first study, a subsample of 35 blocks was selected and rematched by professionals from headquarters. The rematch was done independently of the original match, and then discrepancies between the match and rematch results were adjudicated. Because of this intensive approach to the rematch, we believe the rematch results represent true match status, while differences between the match and rematch results represent the bias in the original match results. Only nonmovers were considered in this study. Also, the study was confined only to within-block rematching, and thus did not formally measure any false nonmatches that may have occurred because the census enumeration was located outside the PES block.

The results for the P sample are given in Table 2 in the form of a cross-tabulation of match statuses as assigned from the original TARO match and the rematch.

We estimate there are about 88 false nonmatches and 18 false matches in the original TARO results, and that $111 = 55 + 56$ cases originally matched or not matched should have been declared to have an indeterminate or unresolved match status. In the normal course of estimation, the unresolved would be treated by missing data procedures (Schenker 1988). The net result is that the observed match rate, i.e., the number matched divided by the number matched plus not matched, is .879 in the original match and .885 in the rematch, and thus that the original match rate is biased downward by about 0.6 percent.

The second evaluation study looked at the extent of matching error for movers. Among the original "not matched," there were 90 persons who reported moving between census-day and the time of the PES. For movers, searching is done at the reported census-day address. As an evaluation of the accuracy of the matching process, we reworked all 90 nonmatched mover cases using more intensive procedures. Eleven new matches were discovered, and as a result, the observed match rate for in-scope movers increased by .058, from .661 to .719. Although, the false nonmatch rate, $11/90 = .122$, for movers is larger than we observed for nonmovers, the movers comprise a relatively small portion of the overall P sample. Correcting the 0.6 percent and 5.8 percent downward bias in match rate for nonmovers and movers has the overall effect of reducing the TARO undercount rate by 0.7 percent.

These calculations ignore the possibility of further new matches that might have been observed had the rematch study extended beyond the bounds of the PES blocks (Thompson,

Whitford and Stoudt 1987). Based on evidence from computer matching across the Los Angeles test site, however, we conclude that geographical assignment was accurate, and that the incremental effect of such additional matches could do no more than to reduce the estimated TARO undercount by a further 0.3 percent.

3. REPORTING CENSUS-DAY ADDRESS

In our new PES design, as in the 1980 design, we attempt to match the P-sample people to the census enumeration at the census-day address. To facilitate the matching, the P-sample interviewer must ask where each household member lived on census day. The interviewer then probes for other addresses where the persons may have lived, including such places as at college or university, on a military base or ship, or at a second home. If the census-day address is reported incorrectly in the P-sample interview, then we may falsely designate the household members as not enumerated in the census, thus biasing upwards the estimated undercount rate.

To study address misreporting, we reinterviewed a subsample of the matched and unmatched cases after the original TARO estimates of undercount had been produced. This followup was six months after the initial PES interview and ten months after census day. Before presenting the results, we mention two limitations on this study. The first is the potential of greater recall error than in the original P-sample interview. Second, any trust created by the census advertising program may have faded, a potentially serious problem in an area with a large number of undocumented immigrants who fear all contacts with the government.

Table 3 describes the composition of the subsample. In most cases, the PES household matches the census household completely ("whole-household matches"). In the category "partial-household matches," some of the PES persons match the census, but others do not. The "whole-household nonmatch with conflicts" category constitutes what we call the "Emerson-Peterson" problem. The census enumerated the "Emersons" at a particular address and the E-sample followup confirmed the census enumeration as correct. However, the P-sample interview showed the "Petersons" as living at the address on census day. These facts are in conflict, and one possible explanation is that the Petersons misreported their census-day address. The "whole household nonmatches without conflicts" category has no apparent contradictions; for example, the census missed the housing unit or listed it as vacant.

Table 3
Post-Production Followup Sample Sizes

Status of Original Match	Number of Households	
	Total in P sample	Rein- terviewed
Whole-Household Match	4,662	50
Partial-Household Match	609	50
Whole-Household Nonmatch with Conflicts	160	64
Whole-Household Nonmatch without Conflicts	357	109

Table 4
Outcome of
Post Production Followup, (Persons) Unweighted

Outcome	Whole-Household NonMatch				Partial-Household Match				Whole- Household Match	
	with Conflict		without Conflict		Non- matched		Matched		#	%
	#	%	#	%	#	%	#	%		
Address Confirmed	64	33	252	73	61	75	138	90	164	99
New Address Given	32	17	46	13	13	16	15	10	1	1
Possible Fabrication	70	36	23	7	2	2	0	0	0	0
Noninterview	27	14	24	7	5	6	0	0	0	0
Total	193	100	345	100	81	100	153	100	165	100

Note: # signifies number of people in the followup subsample.
% signifies percent of column category.

Table 4 gives the results for persons in the sample, with a separate breakout of initially matched v. nonmatched persons in partially matched households. As expected, the rates at which the address was confirmed vary greatly across strata. Virtually all addresses were confirmed for the persons in the whole-household match category, while the lowest rate of confirmation was for the whole-household nonmatch with conflicts category. New addresses were given by 13 to 17 percent of the nonmatched people across each of the three categories. Interestingly, new addresses were reported for ten percent of the matched people within partially matched households, not much less than for the nonmatched people within these households. The newly reported address is unlikely to be correct, unless identical errors were made in the original P sample and census interviews. This variable reporting reinforces our view that followup interviewing months after the original P-sample interview sometimes gives a different response (because of recall error and fear), but not necessarily a more accurate address.

Evidence was gathered on 95 cases that suggest they were possibly fabricated in the original P-sample interview. Most of these cases (70) came from the category of whole household non-matches with conflicts. This problem is discussed further in Section 4. In addition, there were cases where the reinterview was not complete or yielded insufficient information to classify individuals into one of the categories. Some of these, had they been correctly interviewed, may also have reported a new address.

Weighting Table 4 to P-sample totals, we estimate that 3.1 percent of P-sample persons were erroneously reported as nonmovers in the original P-sample interview. For those who moved within the test site, we were able to search for a match at the new address, and we found that one third of those cases were enumerated in the Los Angeles test census. To assess the probable effect of reporting errors, particularly as we view TARO as a test of a national PES, we assume that those people who reported addresses outside the site would have been enumerated at the same rate as those who reported addresses within. Thus, one-third of the 3.1 percent would have been matched and classified as enumerated. Correcting for the reporting error results in a one percent reduction in the estimated undercount.

4. FABRICATION IN THE PES INTERVIEW

In spite of all good efforts to train and control interviewers, a PES interviewer may occasionally fabricate a household in lieu of conducting a proper interview. Fabricated cases will not match to the census. The estimated undercount rate will be inflated to the extent that fabricated cases substitute for people at the address who were actually enumerated.

Our new PES design seeks to control the fabrication rate to low levels. The sample design allows for frequent quality control checks using re-interviews of the interviewers' work. Samples are checked for each interviewer's work from each block several times per week. This close review was not possible in the 1980 PES, where interview assignments were not as highly clustered. We have also improved the training and supervision of the interviewing since 1980. Feedback on performance and retraining is now available to interviewers so that errors will not be repeated.

Two studies shed light on the extent of fabrication in the 1986 TARO. First, extensive quality control checks were performed during data collection for the P sample, both for address listing and for interviewing. The main conclusion from the quality control results is that there was evidence of only a small amount of fabrication. A total of 2070 P-sample interviews were checked by quality control clerks a few days after the original interview to verify the household composition (roster check). Of these, 59 interviews failed the roster check. These cases were examined in detail to determine how many of them were examples of fabrication. This was determined by whether each person in the household, as reported by the original interviewer (not the quality control clerk), matched to the census, which implies that the original interviewer collected valid data for that person. A clone fabrication in the census would be needed to invalidate this assumption. Only 13 of the 59 cases were identified as possible fabrications in that they had, for example, no persons from the original PES roster matching the census. Hence, the estimated fabrication rate for the quality control check is 0.6 percent.

The second source of data on the extent of fabrication is the post-production followup described in Section 3. From the data in Table 4, we estimate that about 1.2 percent of the P-sample people may have been obtained in fabricated interviews. This fabrication rate is about twice as large as provided by the quality control roster check. We believe much of the difference is attributable to one bad interviewer whose work was discovered in the followup interview, but evidently escaped detection by the quality control system. Another part of the difference may be that the followup exaggerates the level of fabrication; that is, landlords and other respondents deny the existence of people who occupy illegally converted housing units or who are present in the country without documentation.

To calculate an upper bound on the effect of fabrication in TARO, we assume the higher fabrication rate, .012, and we assume that if proper interviews had been conducted, the resulting P-sample people would match to the census at the same rate as achieved for the nonfabricated cases, or about .88. This leads to a corrected undercount of about 7.9 percent, about 1.1 percent lower than the original undercount of 9 percent. If we assume the lower fabrication rate, .006, then by similar calculations, the corrected undercount is 8.4 percent, or about .6 percent lower than the original TARO figure. In the summary of TARO errors presented in Table 1, we specified a value of 1 percent, which is about equal to the effect implied by the upper bound.

5. MISSING DATA

In order to measure small coverage errors accurately, the PES data set should be as complete as possible, without a large percentage of missing data. Unfortunately, there was a very large amount of missing data in the 1980 study (Fay *et al.* 1988). A number of changes in the PES design should now lead to lower levels of missingness.

Table 5
PES Missing-Data Rates (%)

Source	1980 PEP		1986 TARO
	April	August	
P Sample			
Noninterview (Household)	4.4	5.3	0.5
Unresolved enumeration status (Person)	4.0	4.4	0.8
Total	8.4	9.7	1.3
Proxy interview (Household)	a	a	3.2
E Sample			
Noninterview (Household)	1.1	1.1	NA
Geocoding indeterminate (Household)	1.6	1.6	NA
Unresolved enumeration status (Person)	2.0	2.0	4.7
Total	4.7	4.7	4.7

^a Percent unknown.
NOTE: NA signifies "not applicable."

First, because of the tight time schedule for CPS interviewing, the initial P-sample interviews in 1980 were conducted during a one-week period. For the new PES, a three-week interviewing period is used, with yet another week if special problems arise. The longer interviewing period decreases the household noninterview rate. Another change that reduces the household noninterview rate is the sample of blocks (rather than list-sample clusters of four housing units as in the CPS). This sample allows the interviewer to visit a housing unit several times (perhaps between visits to the other housing units in the block) without extreme travel costs.

Incomplete followup interviews caused a large portion of the missing P-sample enumeration statuses in the 1980 PES (2.6 percent for April and 2.8 percent for August). We are attempting to diminish this problem by collecting the information needed to declare cases as either enumerated or missed during the initial interview, thereby eliminating the need for followup in most cases. Additionally, improvements in the timing and quality of matching, because of the new automated matcher, will reduce the number of cases requiring followup.

In the new PES design, the P and E samples overlap, and thus most of the information needed to determine E-sample enumeration statuses is gathered early, during initial P-sample interviewing. The use of a block sample, along with improved census geography, also helps reduce the proportion of E-sample cases for which correctness of census geocoding cannot be determined. Finally, improvements have been made in the treatment of missing data (Schenker 1988).

As can be seen in Table 5, the missing-data rates for the P sample in TARO are much lower than those for the 1980 PES. The E-sample total missing-data rate for TARO is equal to that for the 1980 PES, but this was due to an operational error in TARO, and we expect reductions in missing data similar to those for the P sample in the future.

Even though TARO achieved low levels of missing data, it is important to examine what effect the missing data has on the estimated undercount rates. To answer this question, we produced several sets of undercount estimates for TARO derived using alternative treatments of missing data, P-sample proxy interviews, P-sample movers, and certain E-sample unresolved cases. See Schenker (1988) for a detailed description of the alternative estimates, which ranged

from a low of 7.8 percent to a high of 9.4 percent. Two of the alternative treatments considered in Schenker (1988) deal with problems discussed elsewhere in our paper; they are the treatment of movers within the test site (Sections 2 and 3) and E-sample resolved cases that may have been fictitious enumerations (Section 6). The effects of these treatments are attributed in Table 1 to sources of error other than missing data, and are the main reason for the difference between the TARO undercount estimate of 9 percent and the lowest alternative estimate of 7.8 percent. When the other treatments discussed in Schenker (1988) are considered, the change in the estimated undercount ranges from -0.3 percent to 0.3 percent. These changes are quite small and it is uncertain in which direction the true effect lies. Hence, we have listed a mean effect of 0.0 percent in Table 1.

6. ERROR IN MEASURING THE ERRONEOUS ENUMERATIONS

To estimate net coverage error, it is necessary to estimate the number of erroneous enumerations (EE) contained within the original census enumeration. EE includes the following distinct categories:

- (i) fabrication in the census, where the census enumerator or respondent creates fictitious people in lieu of conducting a proper interview;
- (ii) census duplicates;
- (iii) persons born after census-day and persons who died before census-day; and
- (iv) persons enumerated in the census with such sparse or incomplete information as to render them unmatchable to the PES.

All of these categories are estimated by way of the E sample. In addition, certain census geographic coding errors are treated as erroneous enumerations; this problem is part of the balancing issue discussed in Section 7.

In the 1980 PES, the E sample was a separate and independent sample of 110,000 census household enumerations. Interviewers revisited the housing units 8 months after census day to verify that the census enumerations were either correct or erroneous. Also, the housing unit was located on a map to see if it was assigned to the correct census geography, and clerks searched the census records to identify duplicates.

We have instituted two important changes in the new E-sample design. First, as already discussed, both the E and P samples will now be based on the same sample of blocks. We have found that overlapping P and E samples reduces geographic assignment errors. Second, most E-sample data will be collected in July, just three months after census-day. The procedures are such that most E-sample people are automatically designated correctly enumerated if they are counted in the P sample in July and are subsequently matched correctly to the person's E-sample enumeration. Unmatched E-sample cases are tagged for a followup interview, occurring only 6 months after census day. The earlier reporting in this new design lowers the missing data rates, reduces reliance upon proxy respondents, and improves the quality of the collected data.

There are four main components of error in the measurement of EE:

- (i) response errors in the E-sample interview (this is the P-sample interview for most cases and the followup interview for all other cases), or mis-coding of responses by the processing staff;
- (ii) error committed by an interviewer or by staff in assigning the correct geographic code to an E-sample person;
- (iii) error in conducting the search for duplicates; and
- (iv) mistakes made in classifying an E-sample case as having insufficient information for matching.

In addition, there are errors due to non-response in the E-sample interview, as discussed in Section 5, and sampling error, as discussed in Section 9.

Response errors often relate to the assignment of the status of "fictitious" to an E-sample person. The E-sample interviewer sometimes finds that the current resident of a unit (or another eligible respondent) does not know the people listed in the census. Usually, this is because the current resident moved in after the census and simply does not know who was living there at the time of the census. These E-sample cases should be designated as nonresponse. However, if the census enumerations were fabricated, no respondent will know the "people" reported in the census.

In experimenting with the new design in the TARO, the E-sample interviewers were instructed to determine whether the E-sample enumerations were fictitious and to record the basis for their decisions. Initially, the clerks required very strong evidence before designating an E-sample person as fictitious. It was this data that was used in preparing the first TARO estimates of total population and percent undercount. We realized that the rules for coding were being interpreted too strictly, and later, we had professionals review all E-sample cases coded as "noninterview, respondent does not know" to determine if any should have been coded as "fictitious". Out of 257 such E-sample cases, 118 were coded by the professionals as "fictitious." The corrected information was used to create some alternative TARO estimates (Schenker 1988).

Geographic assignment of census returns was generally thought to be very good in the Los Angeles test site, which was a long-established neighborhood with large well-defined blocks. We have not produced formal measures of the effects of geographic misassignment on the estimated EE, but we believe such error is negligible. In other areas of the U.S., however, the errors could be nonnegligible either because of poor maps, poor or incomplete addresses, or confusion about geographic locations created by new construction.

For example, in contrast with Los Angeles, geographic assignment was a problem in the 1986 Mississippi census returns. There we discovered 2.22 percent of the E sample was duplicated. Of the duplicate cases, 35 percent were located outside the sample block. Although we were able to find many duplicates outside the sample block, we are not convinced we found all of them. This is because searching for duplicates was not designed as a separate activity. We only identified duplicates in the course of other PES operations, and thus probably missed many of them. In the next PES, we will implement a separate activity to search for duplicates.

The census sometimes enumerates people with such sparse information that even if they were correctly interviewed in the P sample, a match to the E sample would not be possible. To compensate for this problem, such E-sample cases should be included in EE so as to estimate the total population properly. This problem is similar to that of geographic balancing discussed in Section 7. The separate E and P samples in the 1980 PES made it very difficult to do this consistently; similar cases were classified as "unmatchable" in the E sample and "matchable" in the P sample, thus creating a bias in the dual system estimator. Because the new PES design uses overlapping P and E samples, we ensure that identical rules are applied, thus eliminating the bias.

In another evaluation of the TARO, and as part of the rematch study discussed earlier (see Section 2), the E-sample cases in a subsample of 35 blocks were reprocessed by professionals from headquarters. As in Section 2, the rematch was independent of the original work, with subsequent adjudication of any discrepancies. Thus, we believe the rematch represents the best possible determination of the true enumeration statuses of the E-sample people, while differences between the original work and the rematch may be regarded as a measure of bias due to error in the original work.

Table 6
Results of Rematch Study: E Sample (Weighted)^a

Original Results	Results of Rematching			
	Correct Enumeration	Erroneous Enumeration	Unresolved	Total
Correct Enumeration	19,153	28	88	19,269
Erroneous Enumeration	41	283	1	325
Unresolved	140	100	223	463
Total	19,334	411	312	20,057

^a Weighting is to E-sample totals.

Results are presented in Table 6. Notice that most of the changes involve cases originally classified as “unresolved.” Many of these cases were those discussed earlier, requiring a subjective decision between “fictitious” and “nonresponse.” Based on these data, we believe that better clerical procedures are needed for coding E-sample cases as fictitious. We are presently working to implement improved procedures in the Census Bureau’s next PES, to be done in conjunction with a 1988 dress rehearsal of the 1990 Census.

From the rematch study, we believe the original rate of EE,

$$\frac{325}{325 + 19,269} = .016$$

should be increased to about

$$\frac{411}{411 + 19,334} = .021.$$

This implies the original TARO undercount should be reduced by about 0.5 percent. The corrected undercount is thus about 8.5 percent.

7. BALANCING GROSS OVERCOUNTS AND UNDERCOUNTS

In order to estimate net undercoverage, the methods and concepts used to measure gross overcount must be consistent with those used to measure gross undercount. We refer to this requirement as “balancing.” We proceed to give an elementary description of how the PES achieves balance.

One way to view this issue is to consider the dual system estimator in the form

$$\hat{N}_{++} = (\hat{N}_{1+}\hat{N}_{+1})/\hat{N}_{11},$$

where

$$\hat{N}_{11} = M,$$

the weighted number of matched P-sample people, and

$$\hat{N}_{+1} = N_p,$$

the weighted number of people in the P sample. All notation is defined in Diffendal (1988).

Since we cannot search all census questionnaires, the observed number matched, M , will be lower than the true number in both systems. To make costs manageable, matching for a given case is restricted to a “search area”, typically the sample block and one or two rings of surrounding blocks.

As a consequence, the term \hat{N}_{11} estimates kN^*_{11} , where $0 \leq k \leq 1$ is the conditional probability that a census enumerated individual is counted in the correct search area and N^*_{11} is the PES estimator of N_{11} that would obtain if it were feasible to conduct the search for matches over the entire population.

To construct a consistent estimator of population size, we must reduce the number counted in the census by the factor k . Because the E-sample search for erroneous enumerations, e.g., duplicates, extends over the search area and we treat as erroneous all enumerations that should not be included in the search area, the term \hat{N}_{1+} estimates $k N^*_{1+}$, where N^*_{1+} is the estimator of N_{1+} that would obtain if it were feasible to conduct the search for erroneous enumerations over the entire population.

Assuming consistent search areas, the DSE becomes a consistent estimator of N_{++} . Note that in this model of the balancing process, we do not estimate the probability k , but instead rely on consistent search areas to eliminate it from the DSE.

Balancing the P sample and the E sample in the 1980 PES was impossible because the samples did not overlap. The CPS (or P-sample) addresses were coded to census geography. The search area was to have been limited to a close neighborhood of the CPS address, but because the CPS addresses were based on 1970 Census geography, they could not be easily assigned 1980 Census geographic codes, and searching extended over a wide area. As the search area expanded for the P sample, the E-sample search area should also have expanded. We believe inconsistencies arose between E-and P-sample search areas, thus creating a bias in the DSE.

In TARO, we performed the two-way match between the P-and E-sample persons within the selected blocks. The geography and search areas were consistent, well-defined, and well-controlled during computer and clerical matching. As a consequence, the problem of balancing did not introduce any important bias into the Los Angeles results.

8. CORRELATION BIAS

For the dual system estimator to be a consistent estimator of the true population size N_{++} , two independence assumptions are needed:

- (i) causal independence,
- (ii) heterogeneous independence.

In addition, autonomous independence is often assumed, but failure of this assumption is known to impart little or no bias to the estimate of total population. (Wolter 1986b and Cowan and Malec 1986).

Causal independence fails when an individual’s capture history in the census alters the probabilities of capture in the PES. The estimator \hat{N}_{++} is downward biased when the odds of capture in the PES are increased as a result of capture in the census, and is upward-biased when the odds of capture in the PES are reduced as a result of capture in the census.

An important bias may exist in the April 1980 PES data because of a failure of causal independence. The failure occurred because respondents may have mistaken the April or March CPS enumerations for the census enumeration.

Table 7
Undercounts (%) for Black and Total Population the 1980, 1960 and 1950
U.S. Censuses, and Differential Undercount Rates

Source	Black	Total	Difference
1950			
PES	3.2	1.4	1.8
DA	9.6	4.4	5.2
1960			
PES	3.8	1.9	1.9
DA	8.3	3.3	5.0
1980			
PES ^a			
Low	1.1	-1.0	2.1
Middle	6.9	1.4	5.5
High	5.7	2.1	3.6
DA	5.9	1.4	4.5

^a The 1980 PES produced 12 sets of estimates. The three presented here are selected from the highest, middle and lowest set as measured by estimated total undercount.

Heterogeneous independence fails when census capture probabilities are different from one individual to another. The resulting bias (called heterogeneity bias or correlation bias) is generally thought to be a downward bias because individuals with a high probability of capture in the census also tend to have a high probability of capture in the PES and, conversely, individuals with a low probability of capture in the census also tend to have a low probability of capture in the PES.

Sekar and Deming (1949) suggested post-stratification to control heterogeneity bias. In practical applications, it is unlikely that this technique is fully effective; there is inevitably some residual heterogeneity of capture probabilities within post-strata.

In the dual-system model, the number of people missed by both systems, N_{22} , is estimated by

$$\hat{N}_{22} = \hat{N}_{12}\hat{N}_{21}/\hat{N}_{11},$$

as in Diffendal (1988), equation(2). Because the dual system estimator may be expressed in the form

$$\hat{N}_{++} = \hat{N}_{11} + \hat{N}_{12} + \hat{N}_{21} + \hat{N}_{22},$$

and because \hat{N}_{11} , \hat{N}_{12} , and \hat{N}_{21} are direct design-based estimators, any bias due to failure of the independence assumptions arises solely in \hat{N}_{22} as an estimator of N_{22} .

We can study the correlation bias in 1980 and previous censuses by comparing \hat{N}_{++} to independent demographic analysis (DA) estimates of total population. Table 7 presents relevant data from recent censuses. If one treats demographic analysis estimates as a standard, these comparisons display total bias in the dual system estimator, including both correlation bias and other sources of error. We believe that the downward bias shown in these estimates is largely attributable to correlation bias. The 1950 PES gave severe underestimates of the population size, of the percent undercount, and of the differential undercount, presumably because of both causal and heterogeneity bias. Note, however, that if 1950 PES data had been used to correct the 1950 census, the differential undercount would have been reduced from 5.2 percentage points to approximately 3.4 percentage points.

The 1960 PES gave similar underestimates of population size, of the percent undercount and of the differential undercount, again presumably because of correlation bias. If the 1960 PES data had been used to correct the 1960 census, the differential undercount would have been reduced from 5.0 to approximately 3.1 percent.

No PES was conducted in 1970. The 1980 PES produced 12 sets of estimated undercounts based on the April and August results and on different sets of assumptions. The DA undercount rates are approximately in the middle of the 12 PES undercount rates. Correlation bias is not as evident here as in 1950 or 1960, largely because of improvements that were made in 1980 to reduce positive causal dependence. We believe the heterogeneity bias is still present but is obscured by other PES errors and by bias due to negative causal dependence.

In the new PES design, we attempt to control the bias due to causal effects by scheduling the PES enumeration after most major census field activities. This approach, contrary to that of the April 1980 PES, will promote causal independence between the census and PES enumerations as much as possible. Further, we are now using field office procedures that will promote causal independence, such as assigning PES interviewers to different areas than they worked (if they worked) in the original census enumeration.

It will be difficult to eliminate the correlation bias due to heterogeneity in future PES's. The only possible avenues include more effective post-stratification and combining the PES and DA data in some way, possibly by controlling for DA sex ratios. See Wolter (1986c) and Choi, Steel and Skinner (1988). We have done some experimentation this decade with alternative post-stratification schemes including using variables such as owner/renter status, census mail-back rate, and marital status. These approaches show some promise. See Diffendal (1988).

TARO yielded observed differential undercounts consistent with expected differentials. In the U.S., census coverage is normally lower for males than females. This result has been consistently observed from the results of demographic analysis. The TARO sex ratios (males per 100 females) are higher than the census ratios for Hispanics and for people who were neither Hispanic nor Asian. The TARO sex-ratios are much higher than census sex-ratios (1.1 to 3.4 more males per females) for the 30-44 year age group. This outcome is consistent with the 1980 national results from demographic analysis. Thus, we believe that the TARO sex ratios are closer to the true sex-ratios, and although correlation bias limits the gain, the PES is still able to measure the differential undercount.

Table 8 presents the two-way table of data for the 1986 TARO, with no post-stratification. The estimate of the number missed by both systems,

$$\hat{N}_{22} = 5,870,$$

is approximately the same order of magnitude as census substitutions 5,259 and erroneous enumerations 6,426. Approximately one-eighth of the estimated census misses, $\hat{N}_{12} + \hat{N}_{22} = 44,373$, are attributable to the (2,2) cell. Thus, most of the measured undercount arises from direct survey estimation, not from the dual-system model.

To illustrate the effect of correlation bias, consider doubling the size of the (2,2) cell. This increases the estimated undercount rate by about 1.4 percent. Based upon analysis of the 1980 PES, Ericksen and Kadane (1985) suggest multiplying the (2,2) cell by 2.7, thus increasing the estimated undercount by 2.3 percent.

We have other information that sheds light upon the problem of correlation bias. Three anthropologists worked for the Census Bureau as participant or systematic observers in the Los Angeles test. Their observations do not provide direct measurements of correlation bias, but rather they provide insights into the degree to which the census and PES are missing the

Table 8
Dual-System Estimates for 1986 Los Angeles Test Census

		PES		Total
		Counted	Missed	
Correct Census Enumerations ^a	Counted	298,204	45,463	343,667
	Missed	38,503	5,870	44,373
	Total	336,707	51,333	388,040

^a Correct Census Enumerations = Total Census Enumerations – Substitutions – Erroneous Enumerations.

same kinds of people. The reports suggest that there are people with very low capture probabilities who tend to be missed by both the census and the PES, and thus that an important downward bias may be present in TARO data. See Hainer *et al.* (1988) and Hines (1988).

Given the data available, we have no exact means of assessing the level of correlation bias in the TARO data. Nevertheless, based upon the work just cited, we speculate that the TARO undercount rate may be too small by 2.3 percent or more.

9. RANDOM ERROR

Sampling error affects the estimates of the number of matches, the number of erroneous enumerations, and the P-sample totals. The census count and the number of substituted census people are based upon the 100 percent census enumeration, and as such are not contaminated by sampling error. The estimated standard deviation for the undercount rate is 0.007. So a 95 percent normal-theory confidence interval for the undercount rate is $.09 \pm 2 (.007) = (.076, .104)$.

Diffendal (1988) presents estimated standard errors for the TARO adjustment factors defined by $Y = \hat{N}_{++}/CEN$ and used a components-of-variance model to smooth the Y , thus reducing the effects of sampling error. In most cases, the smoothing substantially reduced the estimated standard errors, particularly for domains. We believe such smoothing can be used profitably in future PES's.

10. CONCLUSION

After the 1980 Census, the Census Bureau reviewed its coverage measurement program and identified the program's weaknesses. We instituted a research program and a new coverage measurement design aimed at reducing the weaknesses. We have completed major tests of the new PES design this decade and have demonstrated substantial improvements over the 1980 PES.

In this article, we reviewed the results of our research program as reflected in the 1986 TARO. There may never be a perfect PES. However, none of the weaknesses or errors in the new design are so large as to invalidate the PES results. For reasons stated in Section 1, we believe the joint effect of the errors in the coverage measurement in TARO is smaller than the error in the original enumeration in Los Angeles.

One of the main benefits of the TARO is that it enables us to identify new questions and minor unresolved problems that warrant further research. For example, the initial PES interview attempted to gather the information needed to declare a P-sample person as missed in the census. We are now refining the questionnaire design, including additional screening questions to identify movers more accurately. In future PES's, we will also conduct followup interviews for most movers and for nonmover households in the P sample suspected of having misreported mover status. In this way, we believe mover misreporting can be kept to a minimum.

The quality control procedures that are intended to detect and correct fabrication in the PES must continue to be improved and tested. In addition to verifying names on the PES roster, other items shall be verified as part of the quality control check. This should detect any partial fabrication that occurs by obtaining names from mailboxes or landlords, and fabricating the characteristics. We are revising the PES followup forms in order to facilitate the identification of fictitious people.

Our goal for future PES's is to minimize missing data, especially through minimizing the need for followup. However, as more cases are sent to followup, the proportion of failed followup cases will increase. Research is needed on the proper treatment of these cases.

Notwithstanding the good results from TARO, one should exercise appropriate caution before drawing the conclusion that the 1990 PES results will be closer to the truth than will the 1990 original enumeration. The actual level of net undercount in the Los Angeles test was high compared to what would be expected in a national census. Will the size of the errors in a national PES be small enough to produce more accurate population estimates?

We believe that the 1990 Census will contain areas with large undercounts and perhaps large overcounts, even if there is a small net national undercount. Thus, the PES should produce the more accurate population estimates for the areas most difficult to count. Through further polishing of the new PES during the last two years of this decade, it may be possible to produce more accurate population estimates for other, less-difficult-to-count areas too.

We also believe that the errors in the PES will decrease as the undercount decreases. Stable areas with good maps, well-defined addresses, few movers and cooperative respondents will be relatively easy for both the census and the PES. Residual processing errors may produce a threshold of accuracy beyond which the PES may not go, regardless of the true net undercount. We will not know for sure until the 1990 PES is executed. This situation may lead the PES estimates to be more accurate than original census estimates for some areas, with equal or nearly equal accuracy for most other areas. Statistical theory should provide a means to produce a best estimate by combining the results of the original enumeration and the PES.

ACKNOWLEDGEMENTS

The results presented represent the work of many people besides the authors, including Dan Childers, Carol Corby, Gregg Diffendal, Charisse Jeffries, Arona Pistiner, Nathaniel Schenker, Maria Urrutia, and Kirsten West. The authors would also like to thank the three referees for their many useful comments.

REFERENCES

- ANOLIK, I. (1988). The rural post-enumeration survey in east central Mississippi. Statistical Research Division Report, Series RR 88/10. U.S. Bureau of the Census, Washington, D.C.

- CHOI, C.Y., STEEL, D.G., and SKINNER, T.J. (1988). Adjusting the 1986 Australian Census for under-enumeration. *Proceedings of the Census Bureau Fourth Annual Research Conference*. Bureau of the Census, Washington, D.C.
- CITRO, C.F., and COHEN, M.L. (1985). *The Decennial Census: New Directions for Methodology in 1990*. Washington: National Academy Press.
- COWAN, C.D. and MALEC, D. (1986). Capture-recapture models when both sources have clustered observations. *Journal of the American Statistical Association*, 81, 347-353.
- DIFFENDAL, G. (1988). The 1986 test of adjustment related operations in central Los Angeles county. *Survey Methodology*, 14.
- ERICKSEN, E.P., and KADANE, J. (1985). Estimating the population in a census year: 1980 and beyond. *Journal of the American Statistical Association*, 80, 98-114.
- FAY, R.E., PASSEL, J.S., ROBINSON, G., and COWAN, C.D. (1988). The coverage of population in the 1980 Census. Technical report PHC 80-E4. Bureau of the Census, Washington, D.C.
- HAINER, P., HINES, C., MARTIN, E., and SHAPIRO, G. (1988). Research on improving coverage in household surveys. *Proceedings of the Fourth Annual Research Conference*, Bureau of the Census, Washington, D.C.
- HINES, C. (1988). The role of participant observation research in understanding the census undercount. Paper presented at the Population Association of America Annual Meetings, New Orleans, La.
- JARO, M. (1988). Advances in record linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association* (forthcoming).
- MARX, R.W. and SAALFELD, A.J. (1988). Programs for assuring map quality at the Bureau of the Census. *Proceedings of the Bureau of the Census Fourth Annual Research Conference*, Bureau of the Census, Washington, D.C.
- MULRY, M., and SPENCER, B. (1988). Total error in dual system estimates of population size. *Proceedings of the Fourth Annual Research Conference*, Bureau of the Census, Washington, D.C.
- SEKAR, C.C., and DEMING, W.E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, 44, 101-115.
- SCHENKER, N. (1988). Handling missing data in coverage estimation, with application to the 1986 test of adjustment related operations. *Survey Methodology*, 14.
- THOMPSON, J.H., WHITFORD, D., and STOUTDT, D. (1987). Memorandum for Howard Hogan, Subject: Review of 1986 PES Matching, April 21, 1987.
- U.S. BUREAU OF THE CENSUS (1978). *The Current Population Survey: Design and Methodology*, Technical Paper No. 40, Washington, D.C.
- U.S. DEPARTMENT OF COMMERCE (1987). Press notes, Statement by Undersecretary Robert Ortner, October 30, 1987.
- WOLTER, K.M. (1986a). Some coverage error models for census data, *Journal of the American Statistical Association*, 81, 338-346.
- WOLTER, K.M. (1986b). A combined coverage error model for individuals and housing units. Statistical Research Division Report Series RR 86/27, U.S. Bureau of the Census, Washington, D.C.
- WOLTER, K.M. (1986c). Capture-recapture estimation in the presence of a known sex ratio. Statistical Research Division Report Series RR 86/20, U.S. Bureau of the Census, Washington, D.C.

Modeling Matching Error and its Effect on Estimates of Census Coverage Error

PAUL P. BIEMER¹

ABSTRACT

Dual system estimators of census undercount rely heavily on the assumption that persons in the evaluation survey can be accurately linked to the same persons in the census. Mismatches and erroneous non-matches, which are unavoidable, reduce the accuracy of the estimators. Studies have shown that the extent of the error can be so large relative to the size of census coverage error as to render the estimate unusable. In this paper, we propose a model for investigating the effect of matching error on the estimators of census undercount and illustrate its use for the 1990 census undercount evaluation program. The mean square error of the dual system estimator is derived under the proposed model and the components of MSE arising from matching error are defined and explained. Under the assumed model, the effect of matching error on the MSE of the estimator of census undercount is investigated. Finally, a methodology for employing the model for the optimal design of matching error evaluation studies will be illustrated and the form of the estimators will be given.

KEY WORDS: Undercount; Dual system estimation; Capture-recapture; Nonsampling error; Processing error.

1. INTRODUCTION

The use of capture-recapture methods for census evaluation and the evaluation of birth-death registration was first suggested by Sekar and Deming (1949). For estimating census coverage error, the method involves matching persons from a sample survey of the population to the census in order to determine the number of individuals which were enumerated in both the sample survey and the census. There are a number of difficulties which may occur in the capture-recapture method to cause substantial biases in an estimate of the total population size, N (see for example Burnham *et al.* 1987 and Wolter 1986). A problem which occurs quite often in applications of the procedure is the failure to accurately match persons from the sample survey to the census. Seltzer and Adlakta (1974) demonstrated that matching error can result in relative biases as large as 33% and may be positive or negative depending upon whether false nonmatches or false matches predominate (see also Scheuren and Oh 1985). Wolter (1983) notes that suspected matching errors in the 1980 Post Enumeration Program were a part of the reason not to adjust the 1980 U.S. Census.

This paper provides a basic framework for evaluating the matching error in capture-recapture studies (particularly for applications to human populations) and for assessing the impact of the errors on the accuracy of the estimate of N . To provide a simple and familiar basis for the discussion of matching error, we shall adopt the original Sekar-Deming capture-recapture model. Extensions of the Sekar-Deming technique are given in Marks, Seltzer and Krotki (1974), and Wolter (1986).

¹ Paul P. Biemer, Head, Department of Experimental Statistics, Director, University Statistics Center, New Mexico State University, Las Cruces, New Mexico, United States.

Consider a population U and let N denote the size of U . A census is conducted and N_c persons are counted. We wish to estimate $N - N_c$ (referred to as the coverage error of the census) which is equivalent to estimating N . A post enumeration survey (PES) is conducted which employs the same reference period as the census. We assume that: (a) both the census and the PES contain no spurious events (i.e., duplications, fabrications, out-of-scope persons or unidentifiable persons) or that the number of such events can be accurately estimated and subtracted from N_c ; and (b) the event of being counted in the census is independent of the event of being counted in the PES.

The PES persons are matched to the census in order to determine the number of PES persons who were also counted in the census. Let x_{11} denote the design unbiased estimator of the total number of persons in both the PES and the census populations and let N_p denote the design unbiased estimator of the PES population size. The Sekar-Deming estimator (more recently referred to as the dual system estimator or DSE) of N is

$$\hat{N} = \frac{N_p N_c}{x_{11}} \quad (1)$$

As we shall see, \hat{N} is subject to two sources of error: sampling error and nonsampling error. Although there may be several sources of nonsampling error, the source of the error of concern here is matching error; i.e., the misclassification of PES persons as enumerated in the census (false positive errors) or not enumerated in the census (false negative errors).

Using Taylor series expansions, general forms for the moments of \hat{N} can be derived. It can be shown that, to terms of order $1/n$, where n is the PES sample size,

$$\text{Bias } (\hat{N}) \doteq -N[\text{Relbias } (\hat{p}_{11}) - \text{Relvar } (\hat{p}_{11})] \quad (2)$$

$$\times [1 + \text{Relbias } (\hat{p}_{11})]^{-1}$$

and

$$\text{Var } (\hat{N}) \doteq N^2 \text{Relvar } (\hat{p}_{11}) [1 + \text{Relbias } (\hat{p}_{11})]^{-2} \quad (3)$$

where $\hat{p}_{11} = x_{11}/N_p$ is an estimator of p_{11} , the true proportion of the PES population falling in the census population; $\text{Relbias } (\hat{p}_{11}) = \text{Bias } (\hat{p}_{11})/p_{11}$; and $\text{Relvar } (\hat{p}_{11}) = \text{Var } (\hat{p}_{11}) \times E^{-2} (\hat{p}_{11})$. Here we have assumed that N_c , the census counts, has a variance of zero. This is a simplification since, as we mentioned, an estimate of the census spurious events may have been subtracted from the census count to obtain N_c and this correction may be subject to sampling and other errors. Nevertheless, the assumption is consistent with our emphasis in this paper on matching error and its effect on \hat{N} . The last section discusses an extension of the methodology which allows error in the estimator N_c .

From (2) and (3) we note that the total mean square error (MSE) of \hat{N} depends upon the total MSE of \hat{p}_{11} . In the following section, we consider some models for evaluating the effects of matching error on \hat{p}_{11} . Letting j ($j = 1, \dots, n$) be the index for the j^{th} individual in the PES sample, we define α_j as the probability that individual j is misclassified in the matching process and consider alternative assumptions regarding the probabilities α_j .

2. MATCHING ERROR MODELS

2.1 Uncorrelated Matching Error

Assume:

1. The event {unit j is misclassified} is independent of the event {unit j' is misclassified} for all $j \neq j'$.
2. $\alpha_j = \theta$ if unit j is truly in the census, referred to as the probability of a false negative error, and $\alpha_j = \phi$ if unit j is truly not in the census, referred to as the probability of a false positive error.

To fix the ideas, we assume simple random sampling for the PES and that n is small relative to N , then

$$E(\hat{p}_{11}) = p_{11}(1-\theta) + (1-p_{11})\phi, \quad (4)$$

$$\text{Bias}(\hat{p}_{11}) = -p_{11}\theta + (1-p_{11})\phi \quad (5)$$

$$\begin{aligned} \text{Var}(\hat{p}_{11}) &= n^{-1} E(\hat{p}_{11}) (1-E(\hat{p}_{11})) \\ &= n^{-1} (SV + SMV), \end{aligned} \quad (6)$$

where SV , denoting *sampling variance*, is given by

$$SV = p_{11}(1-p_{11})(1-\theta-\phi)^2 \quad (7)$$

and where SMV , denoting *simple matching variance*, is given by

$$SMV = p_{11}\theta(1-\theta) + (1-p_{11})\phi(1-\phi) \quad (8)$$

(proof in the appendix).

Readers familiar with the Hansen, Hurwitz, and Pritzker (1964) response error model will recognize the correspondence of their simple response variance and SMV in this model. Hansen, *et al.* define a measure I , referred to as the "index of [response] inconsistency," to be the ratio of the simple response variance to the total variance of a single response, i.e., the proportion of variance which is response variance. For survey responses, I is an indicator of the response reliability of the survey information. An analogous measure can be obtained for matching error to indicate the effect on the variance of \hat{p}_{11} of matching unreliability. This measure, denoted by I_M , is given by

$$I_M = \frac{SMV}{SV + SMV}. \quad (9)$$

For some applications, assumptions (1) and (2) may be too restrictive. The independence assumption (1) is violated, for example, when unit B in the PES is erroneously matched to unit A in the census causing the correct match, unit A in the PES, to be erroneously classified as a nonmatch. Since this implies that the errors for units A and B are negatively correlated, the consequence is that $\text{Var}(\hat{p}_{11})$ will be smaller than given by (6). However, $E(\hat{p}_{11})$ is not affected by correlated errors. Another form of correlated matching error arises when matching is performed by clerks who may vary in their tendencies to commit false positive and false negative errors. The next section provides a model that describes these errors.

Assumption 2 specifies that the misclassification probabilities α_j are homogeneous across the PES population. This too may be a simplification since some individuals, perhaps the majority, may be classified with relatively little risk of error while other individuals are more difficult to match. Basically, matching problems arise from inaccurate or incomplete information about the characteristics of each individual in either or both systems. Therefore, if the PES sample can be post-stratified on the basis of the completeness of the information to be used for matching, the assumption may hold (at least approximately) within each stratum. The overall matching error rate is thus an aggregation of the individual stratum error rates. The last subsection explores this model.

Finally, the assumption of simple random sampling greatly reduces the complexity of the formula for $\text{Var}(\hat{p}_{11})$. Since PES samples are complex samples, the assumption is a simplification, yet it still provides useful formulas for: (a) identifying which components of matching error are likely to have the greatest impact on the total MSE of \hat{N} ; and (b) allocating resources for and designing matching error evaluation studies. In many situations, an adjustment of SV by a "design effect" constant will account for most of the effect of complex sampling on $\text{Var}(\hat{p}_{11})$. Further, $E(\hat{p}_{11})$ is essentially unaffected by more complex forms of sampling than simple random sampling as long as \hat{p}_{11} is appropriately weighted. Thus, the form of $B(\hat{p}_{11})$ does not depend upon this assumption.

2.2 Modeling Clerical Error

Suppose the PES is matched clerically to the census using k clerks. Let m_i denote the number of PES individuals classified by clerk i , $i = 1, \dots, k$. Let the double index (i, j) denote the j^{th} individual in the i^{th} clerk's assignment.

Assume:

1. The event {unit (i, j) is misclassified} and the event {unit (i', j') is misclassified} are independent when $i \neq i'$ and conditionally independent given clerk i for $i = i'$; $j \neq j'$; $i = 1, \dots, k$; $j, j' = 1, \dots, m_i$.
2. $\alpha_{ij} = \theta_i$ if individual (i, j) is truly in the census, and $= \phi_i$ if individual (i, j) is truly not in the census.
3. $E(\theta_i) = \theta$; $E(\phi_i) = \phi$; $\text{Var}(\phi_i) = \sigma_\phi^2$; $V(\phi_i) = \sigma_\phi^2$; and $\text{Cov}(\theta_i, \phi_i) = \sigma_{\phi\theta}$.

For the subset of individuals in the i^{th} clerk's assignment, 1 and 2 are analogous to assumptions 1 and 2 for the model of the last section. Assumption 3 specifies that clerk matching error probabilities are independent and identically distributed random variables. This assumption is analogous to the assumptions made for interviewer errors in interviewer effect models (see for example Kish 1962, Hartley and Rao 1978 and Biemer and Stokes 1985). The assumption is appropriate if our interest lies in estimating the parameters of a much larger pool of clerks of which the k PES clerks are a representative sample.

It is shown in the appendix that, assuming simple random sampling, $E(\hat{p}_{11})$ is still given by (4). The general formula for $\text{Var}(\hat{p}_{11})$ is given by (A.3) in the appendix; however, a useful simplification results if we can assume that the assignment sizes m_i are approximately equal to m , the average size, and that each clerk's assignment has the same expected number of matches (i.e., clerk assignments are interpenetrated). Then

$$\text{Var}(\hat{p}_{11}) = \frac{1}{n} (SV + SMV) + \frac{m-1}{m} \frac{1}{k} CC \quad (11)$$

where CC , denoting the *correlated component of matching variance*, is

$$CC = p_{11}^2 \sigma_\theta^2 + (1-p_{11})^2 \sigma_\phi^2 - 2p_{11}(1-p_{11}) \sigma_{\phi\theta} \quad (12)$$

and SV , SMV are given by (7) and (8), respectively.

Note that CC is a consequence of the between clerk variability of the misclassification probabilities θ_i and ϕ_i . Further, by noting that CC is the variance of $-p_{11} \theta_i + (1-p_{11}) \phi_i$ and the similarity of these terms with (5), we see that CC is the variance of the *net* biases among clerks. This latter fact proves that CC must be positive. Therefore, the effect of clerk variance is to increase the variance of \hat{p}_{11} .

Borrowing again from the response variance literature, we can define a parameter ρ_M which is analogous to the intra-interviewer correlation coefficient, ρ , defined by Kish (1962). We shall refer to ρ_M as the intra-clerk correlation since it is the correlation between the match classifications of any two units in the same clerk assignment. Under the model,

$$\rho_M = \frac{CC}{SV + SMV}$$

is the ratio of the correlated component of variance to the total variance associated with a single classification. It may be interpreted as the degree to which clerks "influence" the match rates within their assignments. Now, an alternative formula for $\text{Var}(\hat{p}_{11})$ which is equivalent to (11) is

$$\text{Var}(\hat{p}_{11}) = \frac{SV + SMV}{n} [1 + (m-1)\rho_M] \quad (13)$$

2.3 Post-stratification

Both the model for uncorrelated error and the model for clerical error assume (essentially) that individuals in the PES sample do not differ in the degree of difficulty of determining their true match classification (assumption 2 for both models). For example, for the clerical error model, the misclassification probability vector (θ_i, ϕ_i) is the same for all units in the i^{th} clerk's assignment. In reality, however, some individuals are much more difficult to classify than others depending upon such factors as the completeness of the matching information, whether a mover or non-mover, whether in single family home or apartment, etc.

A simple approach for modeling this situation is to stratify PES sample according to some variable, say Z , which is correlated with the misclassification probabilities α_j . The variable Z may be an indicator of the completeness of the information, the type of unit, etc.

Suppose there are L such strata indexed by h . Let (i, h, j) denote the j^{th} unit in the h^{th} stratum in the i^{th} clerks assignment where $i = 1, \dots, k$; $h = 1, \dots, L$, $j = 0, \dots, m_{ih}$; and m_{ih} is the number of units in stratum h for the i^{th} clerk. We shall again assume (1) as for the clerical error model; however, in addition assume:

2. $\alpha_{ihj} = \theta_{ih}$ if individual (i, h, j) is truly in the census.
 $= \phi_{ih}$ if individual (i, h, j) is truly not in the census.

3. $E(\theta_{ih}) = \theta_h$; $E(\phi_{ih}) = \phi_h$
 $\text{Var}(\theta_{ih}) = \sigma_{\theta h}^2$; $\text{Var}(\phi_{ih}) = \sigma_{\phi h}^2$;
 $\text{Cov}(\theta_{ih'}, \theta_{ih}) = \sigma_{\theta h}$ if $h = h'$
 $= 0$ if $h \neq h'$

Under these assumptions, we have $\text{Bias}(\hat{p}_{11}) = \Sigma \pi_h \text{Bias}(\hat{p}_{11h})$ and $\text{Var}(\hat{p}_{11}) = \Sigma \pi_h^2 \text{Var}(\hat{p}_{11h}) + \Sigma \pi_h [E(\hat{p}_{11h}) - E(\hat{p}_{11})]^2$ where $\text{Bias}(\hat{p}_{11h})$, $E(\hat{p}_{11h})$, and $\text{Var}(\hat{p}_{11h})$ are given by (5), (4), and (6), respectively, indexing the clerk error parameters and p_{11} by h and where $\pi_h = E(n_h/n)$, the proportion of the population in the h^{th} stratum.

3. DEMONSTRATION OF THE EFFECT ON TOTAL ERROR

The models of the previous section can be useful for demonstrating the effect of matching error on the total mean square error of \hat{N} and \hat{p}_{11} . In the illustrations that follow, we shall assume values of the model parameters which are typical given our experience and which are consistent with current 1990 PES design parameters.

In the PES, estimates of N will be made for a number of census strata. We assume that the desired coefficient of variation of the estimates is 1%. Matching will be conducted in a number of processing sites by teams of clerks. (More details on the matching operation are given in the next section). To illustrate the effect of matching error on the DSE, we consider a "typical" PES stratum. For this stratum, let $p_{11} = .85$ and k , the number of matching clerks in one processing site, be 10. In our analysis, we considered values of θ which varied from 0 to .10 and a number of typical values for the ratio $\gamma = \theta/\phi$, i.e., the ratio of the probability of false negatives to the probability of false positives. Little information exists which would indicate the typical range of ρ_M since no study has ever measured ρ_M for matching error. However, if we assume that the clerk error probabilities θ_i and ϕ_i follow a unimodal beta-distribution and are uncorrelated, we can obtain a maximum value for ρ_M corresponding to given values of the expected error probabilities θ and ϕ . Algebraically, the maximum value of ρ_M is given by

$$\rho_M^* = CC^* / (SMV + SV) \quad (14)$$

where $CC^* = p_{11}^2 \theta^2 (1-\theta) / (1+\theta) + (1-p_{11})^2 \phi^2 (1-\phi) / (1+\phi)$ (see Johnson and Kotz 1970, for the underlying theory). If θ_i and ϕ_i are positively correlated, then the assumption of zero correlation further exaggerates the effect of CC . Thus, the illustrations which follow indicate the maximum impact of matching variance on the estimates.

To illustrate the maximum effect of correlated variance on the precision of \hat{p}_{11} , the coefficient of variation of \hat{p}_{11} , denoted by $CV(\hat{p}_{11})$, was graphed as a function of θ for various values of γ . For these calculations, ρ_M^* was substituted for ρ_M in (13). The range of θ was $0 \leq \theta \leq .10$ and γ was $.5 \leq \gamma \leq 5$; i.e., $\phi = .2\theta$ to $\phi = 2\theta$. This range of values of γ seems reasonable since, typically, ϕ is smaller than θ . Figure 1 shows the function for $\gamma = 1$. There was no discernible difference for other values of γ in the range of interest. Thus, it appears that the size of ϕ has negligible effect on $CV(\hat{p}_{11})$. In fact, we see from the expression for CC^* that when $p_{11} = .85$, no more than 3% of the correlated variance is contributed by the variance of ϕ_i even when ϕ is the same size as θ . Figure 1 also suggest that $CV(\hat{p}_{11})$ may be increased two-fold to 2% for values of θ as small as 5%.

In Figure 2, the relative bias of \hat{p}_{11} , denoted by $RB(\hat{p}_{11})$ is illustrated for the same range of θ ; i.e., $0 \leq \theta \leq .1$, and γ ; i.e., $.5 \leq \gamma \leq 5$. The graph clearly indicates that bias is smaller for smaller values of γ . In fact, the bias is zero when $\gamma = (1-p_{11})/p_{11}$ or .18 assuming $p_{11} = .85$ as in this example. For θ as small as 5%, the relative bias is between -2% and -4%, depending upon the size of γ . Comparing this with the maximum increase in $CV(\hat{p}_{11})$ of one percentage point, we see that bias has the potential to be much more serious than correlated variance.

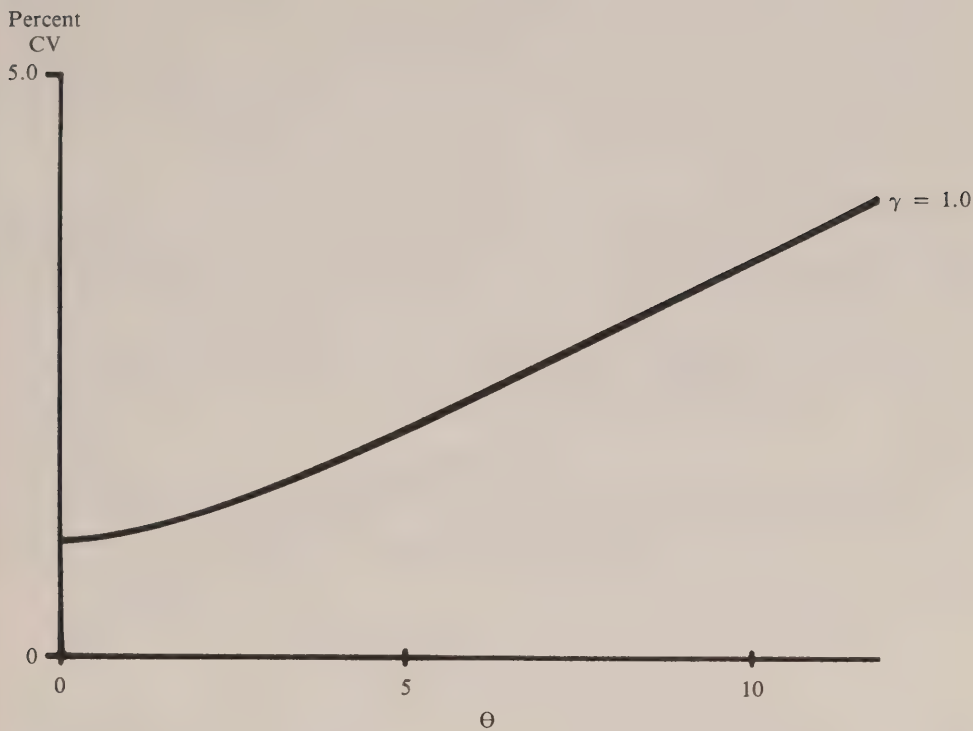


Figure 1. Coefficient of Variation of \hat{p}_{11} as a Function of Θ for $\gamma = 1$

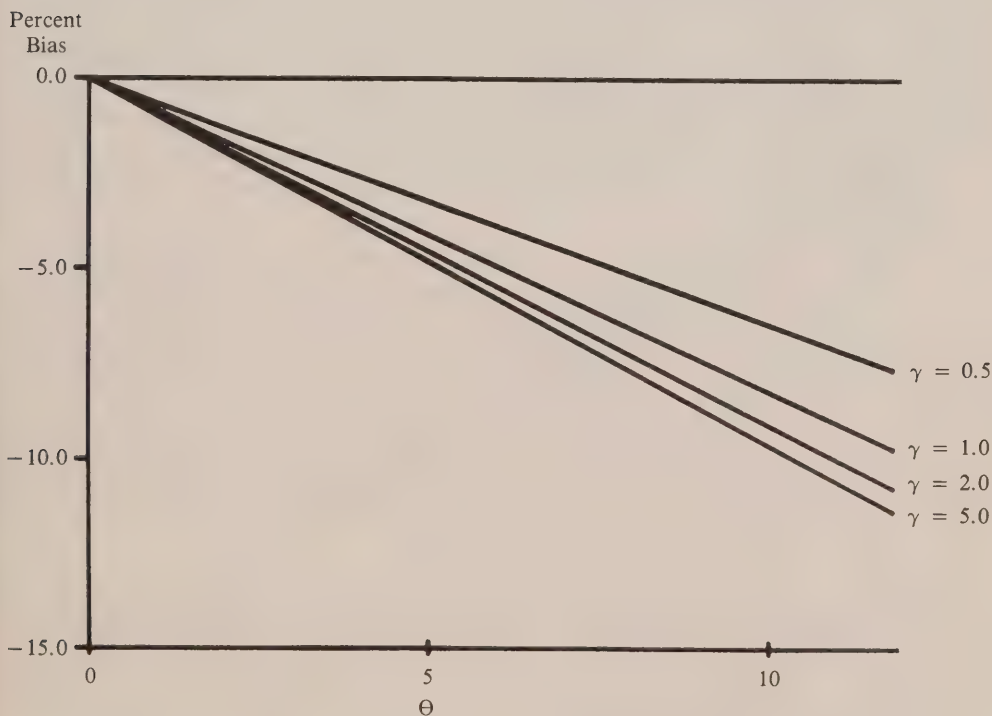


Figure 2. Relative Bias of \hat{p}_{11} as a Function of Θ for Selected Values of γ

To indicate the potential effects of matching error on \hat{N} , the increase in total error as a function of θ and for selected values of γ was computed. Let $M(\theta;\gamma)$, $V(\theta;\gamma)$, and $B(\theta;\gamma)$ denote the mean square error, variance, and bias, respectively of \hat{N} for given values of θ and γ . $M(0;\gamma)$ is the mean square error of \hat{N} without matching error (i.e. $\theta = \phi = 0$) and thus $M(0;\gamma)^{1/2}$ is approximately the standard error of \hat{N} . Define $RM(\theta;\gamma) = (M(\theta;\gamma)/M(0;\gamma) - 1)^{1/2}$; $RV(\theta;\gamma) = (V(\theta;\gamma)/M(0;\gamma) - 1)^{1/2}$; and $RB(\theta;\gamma) = (B^2(\theta;\gamma)/M(0;\gamma))^{1/2}$.

Thus, $RM(\theta;\gamma)$ is the square root of the increase in the total mean square error of \hat{N} for given θ and γ relative to the root MSE of \hat{N} with no matching error. $RV(\theta;\gamma)$ is the contribution of this increase due to matching variance while $RB(\theta;\gamma)$ is the contribution due to matching bias. Hence, we have $RM(\theta;\gamma)^2 = RV(\theta;\gamma)^2 + RB(\theta;\gamma)^2$. Figures 3 and 4 show these functions for two extreme values of γ , $\gamma = .5$ and 5 , respectively, and for $0 \leq \theta \leq .1$. Again, the maximum value of the correlated variance, CC^* , was used for the variance computations. Thus, the contribution of matching variance to total error is probably substantially exaggerated.

These figures indicate that for these values of θ and γ , most of the error is contributed by bias, although the contribution to variance can be non-trivial. Further, as suggested earlier for Figures 1 and 2, the matching bias dominates the total matching error whenever false negative error dominates over false positive error.

4. ESTIMATION FROM REMATCH STUDIES

Methods for estimating the components of response error in sample surveys have been well documented in the literature (see for example Hansen, Hurwitz and Pritzker 1964, Hansen, Hurwitz and Bershad 1961). The techniques for estimating the components of matching error are essentially the same. For example, to estimate the correlated component of matching variance, CC , the assignments of the clerks must be "interpenetrated." This procedure, which is described in detail in Kish (1962), randomizes the assignment of PES cases to clerks so that each clerk's assignment has the same expected number of matched persons. Then, an estimator of CC is formed by the difference between the between clerks and within clerks mean squares from the analysis of variance of clerks. For more details of this procedure, refer to Bureau of the Census (1985).

In this section, the focus is on the analysis of data from rematch studies, the most commonly used method for evaluating matching error. There are two types of rematch studies. One attempts to replicate the original match operation for a sample of cases using the same procedures, training, match rules, etc. This type of rematch has the objective of estimating SMV, the simple matching variance or, equivalently, I_M , the index of match inconsistency. The second type of rematch aims at obtaining the most correct match possible and, therefore, uses more extensive procedures, highly qualified and expert clerks, and adjudication, i.e., resolving disagreements among the original and rematch classifications by a third, expert matcher. This type of rematch has the objective of estimating the matching bias. Further, as we will see, an estimate of SMV is also possible from these data.

The (unweighted) data collected in a rematch study can be displayed as in Table 1. Assume that the rematch sample is a simple random sample of r persons from the PES. Further we may assume either the uncorrelated error model or the clerical error model of the last section for both the match and rematch. Let μ_t ($t = a, b, c, d$) denote the mean observed proportion of the cell corresponding to t in Table 1.

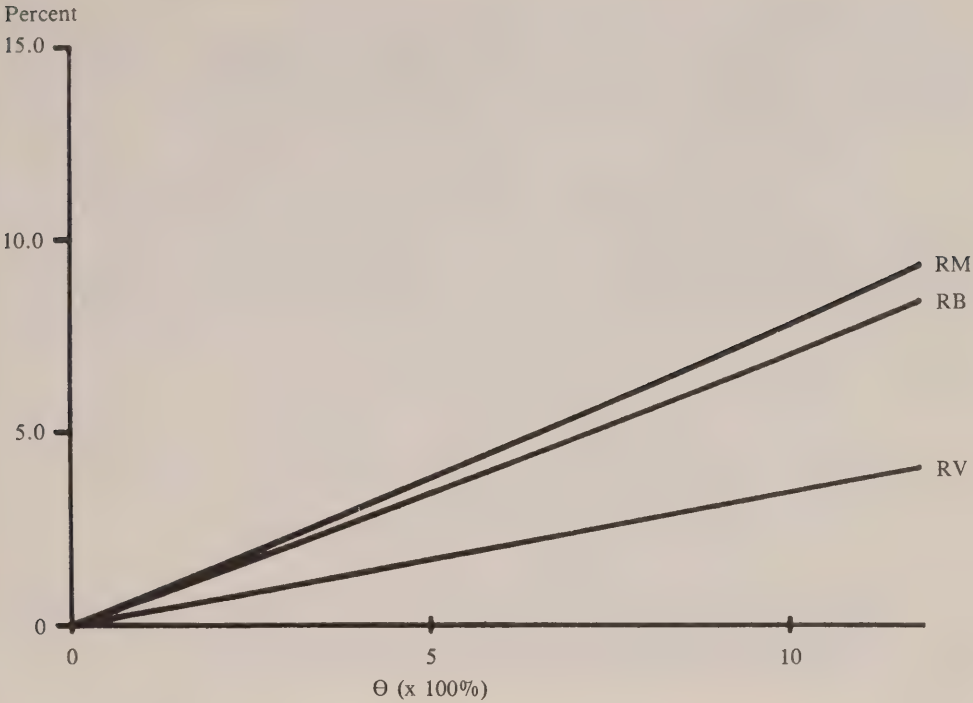


Figure 3. $RM(\Theta; \gamma)$, $RV(\Theta; \gamma)$, and $RB(\Theta; \gamma)$, as a Function of Θ for $\gamma = .5$

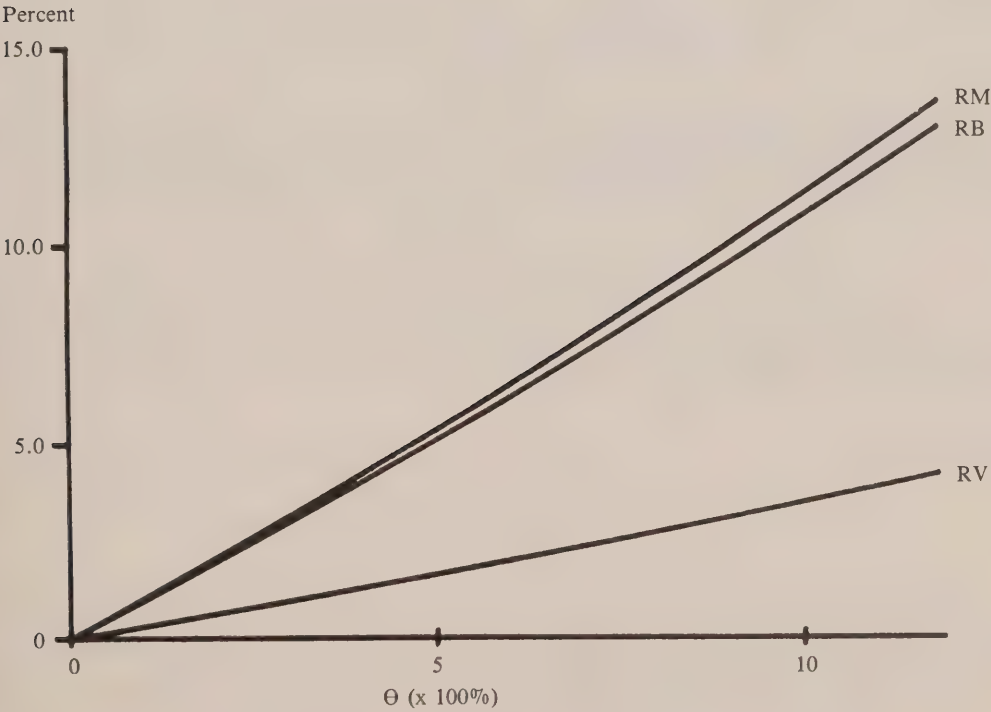


Figure 4. $RM(\Theta; \gamma)$, $RV(\Theta; \gamma)$, and $RB(\Theta; \gamma)$, as a Function of Θ for $\gamma = 5$

Table 1
Rematch Study Data

Original Classification	Rematch Classification	
	Matched	Not Matched
Matched	a	b
Not Matched	c	d

Then

$$\mu_a = p_{11} (1 - \theta_A) (1 - \theta_B) + (1 - p_{11}) \phi_A \phi_B \quad (15)$$

$$\mu_b = p_{11} (1 - \theta_A) \theta_B + (1 - p_{11}) \phi_A (1 - \phi_B) \quad (16)$$

$$\mu_c = p_{11} \theta_A (1 - \theta_B) + (1 - p_{11}) (1 - \phi_A) \phi_B \quad (17)$$

$$\mu_d = p_{11} \theta_A \theta_B + (1 - p_{11}) (1 - \phi_A) (1 - \phi_B) \quad (18)$$

where the index A denotes original match and B denotes the rematch.

Define

$$\mu_A = E\left(\frac{a+b}{r}\right) = p_{11} (1 - \theta_A) + (1 - p_{11}) \phi_A \quad (19)$$

and

$$\mu_B = E\left(\frac{a+c}{r}\right) = p_{11} (1 - \theta_B) + (1 - p_{11}) \phi_B. \quad (20)$$

Note that μ_A and μ_B are expected values of the estimates of p_{11} based upon the original and the rematch classifications, respectively. The difference of these two estimates of p_{11} , i.e., $(b - c) / r$ is referred to as the *net difference rate* (NDR). Its expected value is

$$E(NDR) = \mu_A - \mu_B = -p_{11}(\theta_A - \theta_B) + (1 - p_{11}) (\phi_A - \phi_B). \quad (21)$$

Finally, the proportion of the r sample individuals having rematch classifications which disagree with the original match classification is $(b + c) / r$, referred to as the *gross difference rate* (GDR). Its expected value is

$$\begin{aligned} E(GDR) &= \mu_b + \mu_c \\ &= p_{11} [\theta_A (1 - \theta_B) + (1 - \theta_A) \theta_B] + (1 - p_{11}) [(1 - \phi_A) \phi_B + \phi_A (1 - \phi_B)]. \end{aligned} \quad (22)$$

We shall now consider the estimation of the components of $\text{Var}(\hat{p}_{11})$ and $\text{Bias}(\hat{p}_{11})$ under three sets of assumptions for the rematch study. In the first case, we assume that the rematch study is conducted under the same general conditions as the original match so that the error parameters associated with both classifications are very nearly the same. For example, the clerks for both operations received the same training, have the same skill level, and use the same

procedures. The second case assumes that the rematch is perfect, i.e., the rematch classification may be considered the true classification. The third case falls somewhere between case 1 and 2. More extensive and improved matching procedures are used in the rematch; however, we are not willing to assume that the rematch classifications are without error. Instead we assume that fewer errors are made in the rematch than in the original match.

Case 1. Same General Conditions for the Match and Rematch

Assume that $\theta_A = \theta_B = \theta$ and $\phi_A = \phi_B = \phi$, i.e., the expected rates of misclassification are the same for both trials. Then, from (21), $E(NDR) = 0$ and no estimate of Bias (\hat{p}_{11}) can be computed from the data. However, from (22) and (8)

$$\frac{1}{2} E(GDR) = SMV \tag{23}$$

Further, an estimator of I_M in (9) is

$$\hat{I}_M = GDR / [2 \hat{p}_{11} (1 - \hat{p}_{11})] \tag{24}$$

where \hat{p}_{11} is the PES estimator of p_{11} as defined for (2). Alternatively, an estimator of $E(\hat{p}_{11})$ can be obtained from Table 1; for example, see the estimators in (19) and (20).

Case 2. Perfect Rematch

Assume that $\theta_B = \phi_B = 0$, i.e., the rematch is conducted without misclassification error. Then, from (21),

$$\begin{aligned} E(NDR) &= -p_{11} \theta_A + (1 - p_{11}) \phi_A \\ &= \text{Bias } (\hat{p}_{11}). \end{aligned} \tag{25}$$

Further, the probability of false negative error, θ_A , is estimated by

$$\hat{\theta} = c / (a + c). \tag{26}$$

and, the probability of false positive error, ϕ_A , is estimated by

$$\hat{\phi} = b / (b + d). \tag{27}$$

An estimator of SMV is

$$\widehat{SMV} = \frac{1}{r} \left(\frac{ac}{a + c} + \frac{bd}{b + d} \right) \tag{28}$$

and, thus, an estimator of I_M is

$$\hat{I}_M = \widehat{SMV} / \hat{p}_{11} (1 - \hat{p}_{11}) \tag{29}$$

where \hat{p}_{11} is an estimator of $E(\hat{p}_{11})$ obtained either from the PES or from Table 1.

Case 3. Rematch Has Smaller Error But is Not Perfect

Assume that $0 < \theta_B < \theta_A$ and $0 < \phi_B < \phi_A$; i.e., the misclassification probabilities for the rematch are smaller than for the original match but are not zero. Then no unbiased estimator of Bias (\hat{p}_{11}) exists. However, $|E(NDR)|$ will be smaller than $|\text{Bias}(\hat{p}_{11})|$ if $\mu_A - p_{11}$ and $\mu_B - p_{11}$ both have the same sign; i.e., the estimator of p_{11} based on the match and the rematch data are biased in the same direction. Thus, under these conditions, $|NDR|$ is a lower bound estimator of $|\text{Bias}(\hat{p}_{11})|$.

Further, there is no unbiased estimator of SMV. However, it can be seen from (22) that

$$E(GDR) - 2SMV = p_{11}(\theta_B - \theta_A)(1 - 2\theta_A) + (1 - p_{11})(\phi_B - \phi_A)(1 - 2\phi_A).$$

Thus, whenever θ_A and ϕ_A are both less than .5, which is true in most practical applications, we have

$$E(GDR) < 2SMV$$

and \hat{I}_M defined in (24) will underestimate I_M .

5. APPLICATION TO THE 1990 CENSUS

In the 1990 Census, the PES sample will consist of about 5000 "blocks" or groups of about 30 contiguous housing units and attempts will be made to match each person in every block to the census. The variables used for matching will include Name, Address, Relation to Head of Household, Sex, Birthdate, Marital Status, Race, and Hispanic Origin. The matching process will involve four separate stages as follows:

- Stage 1. A computer match operation using the Fellegi and Sunter (1969) technique. Each PES person will be classified as either matched to the census, not matched, or possibly matched (i.e., requiring clerical review) by computer.
- Stage 2. A first clerical review to correct any mismatches or erroneous non-matches made by the computer. In addition, a standardized set of matching rules will be applied to each possible match. Thus, each PES person will be classified as either a match, a non-match, a possible match or an unresolved case.
- Stage 3. A second clerical review to reconsider, by applying greater human judgment, the classification made at the two earlier stages. The clerks for this stage, referred to as the special matching group (SMG), may also decide that for some households further field follow-up is required.
- Stage 4. An "after field follow-up" review. Cases are reconsidered on the basis of any additional information obtained in the follow-up. The final classification codes are matched (enumerated), not matched (not enumerated) or unresolved (match status to be imputed in the final processing stage).

The procedures for imputing "matched" or "not matched" for unresolved cases are described in Schenker (1987). These cases which account for about 1% of the PES sample are not included in the tables which follow since the imputed match statuses of the unresolved cases were not available for this test. Nevertheless, imputation error can be an important source of matching error — one which poses special problems for the evaluation. For example, it is likely that some of the PES unresolved cases will also be unresolved in the rematch and no direct estimate of misclassification error can be computed for these cases. In the test described below, 83% of the unresolved PES cases remained unresolved in the rematch. Conversely, 41% of the cases which were unresolved in the rematch, were resolved in the PES match. If one assumes that imputations for those cases which were unresolved in the rematch are erroneous, an upper bound on the imputation error can be obtained. Likewise, a lower bound can be obtained by assuming all these imputations are correct. However, unless the proportion of imputations is very small, this "worst-case, best-case" analysis may yield bounds which are too wide to be useful.

In 1986, a pretest of these PES matching procedures was conducted in Los Angeles. A sample of about 4000 persons were matched to the Los Angeles test census and then rematched by census professional staff to evaluate matching bias. Special procedures were used in the rematch to ensure a very accurate match classification. Table 2 displays the rates of disagreement among the four stages of matching and the rematch. Note the improvement of the classifications at each higher stage indicated by the decreasing disagreement rate in the rematch column. The data also indicate that few classifications are affected in the "after follow-up" stage (.68% disagreement with stage 3). Further, the GDR for the final stage (relative to the rematch) is very low, less than 1%.

Under the assumption that the rematch process yields the true match status, Table 3 gives the estimates of θ , the probability of false negative error, and ϕ , the probability of false positive error, for each stage of matching. It appears, that for the computer match and the first level clerical match, the false nonmatch rate predominates. However, the opposite is true for the final two stages of matching.

Table 2
Comparison of Disagreement Rates for Stages of Matching (%)

	Stage 2	Stage 3	Stage 4	Rematch
Stage 1	2.9	4.4	4.7	5.5
Stage 2	0	3.3	4.0	4.8
Stage 3	3.3	0	.68	1.6
Stage 4	4.0	.68	0	.87

Table 3
Estimates of θ and ϕ for Stages of Matching

Stage of matching	Estimate of θ (x100%) (false nonmatch rate)	Estimate of ϕ (x100%) (false match rate)
1	6.2	2.3
2	5.1	3.3
3	1.5	2.1
4	.1	.3

Table 4
Results of the Rematch Study (weighted)

Original Match Classification	Rematch Classification	
	Matched	Not Matched
Matched	16690	9
Not Matched	85	2178

Table 5a
Rematch Results For Cases With Agreement On All Four Stages.

Original Match Classification	Rematch Classification	
	Matched	Not Matched
Matched	14458	0
Not Matched	64	1775

Table 5b
Rematch Results For Cases With Disagreement On at Least One Stage.

Original Match Classification	Rematch Classification	
	Matched	Not Matched
Matched	2223	9
Not Matched	21	403

Using the methodology of the previous section, we can estimate Relbias (\hat{p}_{11}), Relbias (\hat{N}), and I_M , the index of match inconsistency. Table 4 gives the results of the rematch study, weighted for the rematch sample probabilities of selection. For this table, the estimate of Relbias (\hat{p}_{11}) is $-.4\%$ and therefore, the estimate of Relbias (\hat{N}) is $.4\%$, computed from (2) assuming a 1% coefficient of variation for \hat{p}_{11} and replacing Relbias (\hat{p}_{11}) by its estimate. I_M is estimated to be $.49\%$ which is in the very low range. The false positive rate is $\hat{\phi} = .004$ and the false negative rate is $\hat{\theta} = .005$.

As mentioned in the second section, the probability of matching error may depend upon the completeness of the PES or census information, among other things. To indicate the extent to which match error rates vary, the rematch sample was partitioned into two subsamples. The first subsample was composed of cases which were classified as “matched” or “not matched” consistently across all stages of matching, i.e., for which all four stages agreed. The remainder of the sample made up the second subsample, i.e., cases for which at least one of the stages disagreed. This division approximates a division based upon completeness of the matching information since most of the cases having no disagreement between stages are those where information is the most complete. The weighted results are shown in tables 5a (complete cases) and 5b (incomplete cases).

For "complete" cases, the false negative rate is .44% while the false positive rate is 0. Thus, none of the cases were erroneously matched although a modest number were erroneously called nonmatches. These data may provide evidence of the greater skill of the rematch staff at finding matches for PES cases. The estimate of I_M is .39%, very low. For "incomplete" cases, the false negative rate is .93% while the false positive rate is 2.18%. The estimate of I_M is 1.1%, still quite low. However, these data indicate a much higher risk of false matches for the "incomplete" cases.

The data from this study indicates that matching error causes a small negative bias ($-.4\%$) in \hat{N} which amounts to an underestimate of approximately one million persons (assuming $N = 250$ million persons). Even for the more difficult cases the bias is only $-.7\%$. It would be interesting to look at certain demographic subgroups of the population — movers, proxy respondents, and apartment dwellers — to see the extent of matching error for these domains. Unfortunately, the information that would allow this analysis is not currently available.

6. SUMMARY

The models and MSE formulas developed in this paper can be useful for evaluating the impact of matching error on estimates of census coverage error. In the context of the 1990 U.S. census matching error bias appears to be the largest and most important component of the $MSE(\hat{N})$. Preliminary studies of the magnitude of matching error bias for the 1990 Census indicate that this component is small, less than one half of one percent. This estimate does not reflect imputation error which affects about 1% of the PES cases. Moreover, estimates of bias depends heavily on the assumption that the rematch process yields the true match classification. More work is needed to check the validity of this assumption.

In the development of the formulas for the total mean square error of \hat{N} , we assumed that N_c was not prone to error. However, in actual practice, an estimate of the numbers of census spurious events (or erroneous enumerations), denote by EE, may be subtracted from N_c . Since this estimator is obtained from a match of a sample of the census units to the PES, EE is also subject to sampling error and matching error. For example, a person may be classified as an erroneous enumeration when they were correctly enumerated (false positive error), or they may be classified as correctly enumerated when they are erroneously enumerated (false negative error). The model and methodology formulated for evaluating the effect of false positive and false negative errors for x_{11} can be easily extended for the estimator of erroneous enumerations. Note that the Taylor approximation formulas for the bias and variance of \hat{N} , (2) and (3), will now contain terms for the bias and variance of EE.

For future research, studies of matching error correlated variance are needed to inform us of the extent to which the clerk variance contributes to the total error of \hat{N} . We suspect that CC^* , the maximum effect of correlated error, substantially over estimates the impact of clerks. Research is also needed from rematch studies to identify the characteristics of persons or households prone to matching error. Perhaps then special efforts could be directed toward these cases. For this objective, the use of logistic models should be explored for predicting the probability a case is misclassified from the various characteristics of the case.

ACKNOWLEDGMENTS

This work was supported though a Joint Statistical Agreement with the U.S. Bureau of the Census. I wish to thank Aref DeJani of the Census Bureau for providing some computer support for the preparation of this paper. Thanks are also due to Bernice Garrett for typing and proof reading of the paper.

APPENDIX

Derivation of the MSE Formulas

Let U denote the population of size N to be enumerated. Let U_c denote the subset of U which is enumerated in the census. Let S denote the PES sample and S_c denote $S \cap U_c$, the set of PES persons enumerated in the census. Denote the n units in S as u_1, \dots, u_n . Define the variables

$$\eta_i = 1 \text{ if } u_i \in S_c$$

$$= 0 \text{ if } u_i \notin S_c$$

and

$$y_i = 1 \text{ if } u_i \text{ classified (by the matching process) in } S_c.$$

$$= 0 \text{ if } u_i \text{ not classified in } S_c.$$

Model for Correlated Error

Assume: (1) y_i is a random variable with $P(y_i = 1 | \eta_i = 0) = \phi$ and $P(y_i = 0 | \eta_i = 1) = \theta$, and (2) y_i and y_j are independent given η_i and η_j for $i \neq j$. Let $E(\cdot | S)$ and $V(\cdot | S)$ denote conditional expectation and variance, respectively, given S . Then, $\hat{p}_{11} = \Sigma y_i / n$ and $E(\hat{p}_{11} | S) = (1 - \theta)\tilde{p}_{11} + \phi(1 - \tilde{p}_{11})$ where $\tilde{p}_{11} = \Sigma \eta_i / n$. Taking expectation with respect to S yields the result in (5).

Further, $V(y_i | \eta_i = 0) = \phi(1 - \phi)$ and $V(y_i | \eta_i = 1) = \theta(1 - \theta)$. Therefore, $V(\hat{p}_{11} | S) = \phi(1 - \phi)(1 - \tilde{p}_{11}) / n + \theta(1 - \theta)\tilde{p}_{11} / n$.

Taking expectation with respect to S yields SMV in (8).

Finally, combining $VE(\hat{p}_{11} | S)$ and $EV(\hat{p}_{11} | S)$ yields the result in (6).

Model for Clerical Error

Let (i, j) denote the j^{th} person in the i^{th} clerk's assignment. Let y_{ij} and η_{ij} be defined in analogy to y_i and η_i . Assume (1) — (3) for the clerical error model. Let E_2 , V_2 , and C_2 denote conditional expectation, variance, and covariance with respect to the clerk error distributions holding the sample of clerks fixed. Let E_1 , V_1 , and C_1 denote the corresponding expectation, variance and covariance with respect to the random selection of the k clerk parameter vectors, as per assumption (3), holding the sample S fixed. Then

$$\begin{aligned} E_1 E_2 (\hat{p}_{11}) &= E_1 \left\{ \sum_i [(1 - \theta_i) \frac{n_{1i}}{n} + \phi_i \frac{n_{0i}}{n}] \right\} \\ &= (1 - \theta)\tilde{p}_{11} + \phi(1 - \tilde{p}_{11}) \end{aligned}$$

where $n_{1i} = \sum_j \eta_{ij}$ and $n_{0i} = \sum_j (1 - \eta_{ij})$. Hence, (4) follows upon taking expectation of (A.1) with respect to S .

Consider the variance of \hat{p}_{11} . We have $\text{Var}(\hat{p}_{11}) = VE(\hat{p}_{11} | S) + EV(\hat{p}_{11} | S)$ where $E(\hat{p}_{11} | S)$ is given by (A.1). Further $n^2 V(\hat{p}_{11} | S) = \sum_i \sum_i V(y_{ij} | S) + \sum_i \sum_{j \neq j'} \text{Cov}(y_{ij}, y_{ij'} | S)$

where $V(y_{ij} | S) = V_2(y_{ij}) + V_1 E_2(y_{ij})$ and $\text{Cov}(y_{ij}, y_{ij'} | S) = C_1 [E_2(y_{ij}), E_2(y_{ij'})]$, the term $E_1 C_2(y_{ij}, y_{ij'})$ being zero. Since $E_2(y_{ij}) = \phi_i$, for $\eta_{ij} = 0$, and $E(y_{ij}) = 1 - \theta_i$ for $\eta_{ij} = 1$, we have $V_1 E_2(y_{ij}) = \sigma_{\phi}^2$, if $\eta_{ij} = 0$, and $= \sigma_{\theta}^2$ if $\eta_{ij} = 1$. Further $V_2(y_{ij}) = \phi_i(1 - \phi_i)$ for $\eta_{ij} = 0$ and $V_2(y_{ij}) = \theta_i(1 - \theta_i)$ for $\eta_{ij} = 1$. Thus

$$\begin{aligned} E_1 V_2(y_{ij}) &= \phi(1 - \phi) - \sigma_{\phi}^2 \text{ if } \eta_{ij} = 0 \\ &= \theta(1 - \theta) - \sigma_{\theta}^2 \text{ if } \eta_{ij} = 1 \end{aligned}$$

Similarly, it can be shown that, for $j \neq j'$,

$$\begin{aligned} C_1 \{E_2(y_{ij}), E_2(y_{ij'})\} &= \sigma_{\theta}^2 \text{ if } (\eta_{ij}, \eta_{ij'}) = (1, 1) \\ &= -\sigma_{\phi\theta} \text{ if } (\eta_{ij}, \eta_{ij'}) = (1, 0) \\ &= \sigma_{\phi}^2 \text{ if } (\eta_{ij}, \eta_{ij'}) = (0, 0). \end{aligned}$$

Therefore,

$$V(\hat{p}_{11} | S) = (\Sigma m_i^2 - n) / n^2 CC + SMV / n. \tag{A.2}$$

Finally, combining (A.1) and (A.2) in the identity

$$\begin{aligned} V(\hat{p}_{11}) &= VE(\hat{p}_{11} | S) + EV(\hat{p}_{11} | S), \text{ we have} \\ V(\hat{p}_{11}) &= 1 / n(SV + SMV) + (\Sigma m_i^2 - n) / n^2 CC. \end{aligned} \tag{A.3}$$

If we further assume that $m_i = m$ for all i we obtain the form in (11).

REFERENCES

BIEMER, P. P., and STOKES, S. L. (1985). Optimal design of interviewer variance experiments in complex surveys. *Journal of American Statistical Association* 80, 158-166.

Bureau of the Census (1985). *Evaluating censuses of population and housing*, Statistical Training Document. United States Bureau of the Census, Washington, D.C.

BURNHAM, K. P., ANDERSON, D. R., WHITE, G. C., BROWNIE, C., and POLLOCK, K. H. (1987). *Design and Analysis Methods for Fish Survival Experiments Based on Release — Recapture*. American Fisheries Society Monograph 5.

FELLEGI, I. P., and SUNTER, A. B. (1969). A Theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

HANSEN, M. H., HURWITZ, W. N., and BERSHAD, M. A. (1961). Measurement errors in censuses and surveys. *Bulletin of the International Statistical Institute*, 38, 359-374.

HANSEN, M. H., HURWITZ, W. N., and PRITZKER, L. (1964). The Estimation and interpretation of cross differences and the simple response variance. In *Contributions to Statistics* (Ed. C. R. Rao), Oxford: Pergamon Press, 111-136.

HARTLEY, H. O., and RAO, J. N. K. Estimation of nonsampling variance components in sample surveys. In *Survey Sampling and Measurement*, (Ed. N.K. Namboodiri), New York: Academic Press.

JOHNSON, N. L., and KOTZ S. (1969). *Continuous Univariate Distributions II*. Boston: Houghton Mifflin.

- KISH, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, 57, 92-115.
- MARKS, E. S., SELTZER, W., and KROTKI, K. J. (1974). *Population Growth Estimation*. New York: Population Council.
- RICHER, W. E. (1958). *Handbook of Computations for Biological Statistics of Fish Populations*. Fisheries Research Board of Canada. Ottawa: Queen's Printer and Controller of Stationery.
- SEBER, G. A. F. (1982). *The Estimation of Animal Abundance and Related Parameters*, (3rd. ed.). New York: MacMillan.
- SEKAR, C. C., and DEMING, W. E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, 44, 101-115.
- SELTZER, W., and ADLAKTA A. (1974). On the effect of errors in the application of the Chandrasekaran — Deming techniques, Reprint 14. Laboratory for Population Statistics, University of North Carolina.
- SCHENKER, N. (1987). Report on missing data in the 1986 test of adjustment related operations. Survey Research Division Report Series, Bureau of the Census, RR-87/09.
- SCHEUREN, F., and OH, H. L. (1985). Fiddling around with nonmatches and mismatches. *Proceedings of the Workshop on Exact Matching Methodologies*. Arlington, Virginia.
- WOLTER, K. M. (1983). Affidavit, Mario Cuomo *et al.* vs. Malcolm Baldrige *et al.* U. S. District Court, Southern District of New York, 80 Civ. 4550.

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 10, No. 2 and onward) of *Survey Methodology* as a guide and note particularly the following points:

1. Layout

- 1.1 Manuscripts should be typed on white bond paper of standard size ($8\frac{1}{2} \times 11$ inch), one side only, entirely double spaced with margins of at least $1\frac{1}{2}$ inches on all sides.
- 1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, etc.
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (e.g., w, ω ; o, O, 0; l, 1).
- 3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, priez d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 10, n° 2) et de noter les points suivants:

1. **Présentation**
 - 1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.
 - 1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés. Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
 - 1.4 Les remerciements doivent paraître à la fin du texte.
 - 1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.
2. **Résumé**

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. Rédaction

- 3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
- 3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(·) et log(·) etc.
- 3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
- 3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
- 3.5 Distinguer clairement les caractères ambigus (comme w, ω; o, O; l, I).
- 3.6 Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.

4. Figures et tableaux

- 4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
- 4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois.)

5. Bibliographie

- 5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.
- 5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

- HANSEN, M. H., HURWITZ, W. N., et PRITZKER, L. (1964). The estimation and interpretation of cross differences and the simple response variance. Dans *Contributions to Statistics* (éd. C. R. Rao), Oxford: Pergamon Press, 111-136.
- HARTLEY, H. O., et RAO, J. N. K. Estimation of nonsampling variance components in sample surveys. Dans *Survey Sampling and Measurement*, (éd. N. K. Namboodiri), New York: Academic Press.
- JOHNSON, N. L., et KOTZ S. (1969). *Continuous Univariate Distributions II*. Boston: Houghton Mifflin.
- KISH, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, 57, 92-115.
- MARKS, E. S., SELTZER, W., et KROTKI, K. J. (1974). *Population Growth Estimation*. New York: Population Council.
- RICKER, W. E. (1958). *Handbook of Computations for Biological Statistics of Fish Populations*. Fisheries Research Board of Canada. Ottawa: Queen's Printer and Controller of Stationery.
- SEBER, G. A. F. (1982). *The Estimation of Animal Abundance and Related Parameters*, (3^e éd.). New York: MacMillan.
- SEKAR, C. C., et DEMING, W. E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, 44, 101-115.
- SELTZER, W., et ADLAKTA A. (1974). On the effect of errors in the application of the Chandrasekaran — Deming techniques, Reprint 14. Laboratory for Population Statistics, University of North Carolina.
- SCHENKER, N. (1987). Report on missing data in the 1986 test of adjustment related operations. Survey Research Division Report Series, Bureau of the Census, RR-87/09.
- SCHUREN, F., et OH, H. L. (1985). Fiddling around with nonmatches and mismatches. *Proceedings of the Workshop on Exact Matching Methodologies*. Arlington, Virginia.
- WOLTER, K. M. (1983). Affidavit, Mario Cuomo et coll. vs. Malcolm Baldrige et coll. U. S. District Court, Southern District of New York, 80 Civ. 4550.

où $n_i = \sum_j \eta_{ij}$ et $n_{0i} = \sum_j (1 - \eta_{ij})$. Donc, (4) est obtenu en calculant l'espérance de

(A.1) par rapport à S .

Considérons la variance de \hat{p}_{11} . On a $\text{Var}(\hat{p}_{11}|S) = \text{VE}(\hat{p}_{11}|S) + \text{EV}(\hat{p}_{11}|S)$ où $\text{E}(\hat{p}_{11}|S)$ provient de (A.1). De plus $n^2 \text{V}(\hat{p}_{11}|S) = \sum_j^i \sum_{j \neq i}^i \text{V}(y_{ij}|S) + \sum_j^i \text{Cov}(y_{ij}, y_{ij'}|S)$ où $\text{V}(y_{ij}|S) = \text{V}_2(y_{ij}) + \text{V}_1 \text{E}_2(y_{ij})$ et $\text{Cov}(y_{ij}, y_{ij'}|S) = \text{C}_1 [\text{E}_2(y_{ij}), \text{E}_2(y_{ij'})] + \text{E}_1 \text{C}_2(y_{ij}, y_{ij'})$ étant égal à zéro. Étant donné que $\text{E}_2(y_{ij}) = \phi_i$ pour $\eta_{ij} = 0$, et que $\text{E}(y_{ij}) = 1 - \theta_i$ pour $\eta_{ij} = 1$, $\text{V}_1 \text{E}_2(y_{ij}) = \sigma_{\theta}^2$, si $\eta_{ij} = 0$, et $\text{V}_1 \text{E}_2(y_{ij}) = \sigma_{\theta}^2$ si $\eta_{ij} = 1$. De plus, $\text{V}_2(y_{ij}) = \phi_i(1 - \phi_i)$ pour $\eta_{ij} = 0$ et $\text{V}_2(y_{ij}) = \theta_i(1 - \theta_i)$ pour $\eta_{ij} = 1$. Ainsi

$$\text{E}_1 \text{V}_2(y_{ij}) = \phi_i(1 - \phi_i) - \sigma_{\theta}^2 \text{ si } \eta_{ij} = 0$$

$$= \theta_i(1 - \theta_i) - \sigma_{\theta}^2 \text{ si } \eta_{ij} = 1$$

De la même façon, on peut démontrer que, pour $j \neq j'$,

$$\text{C}_1 \{ \text{E}_2(y_{ij}), \text{E}_2(y_{ij'}) \} = \sigma_{\theta}^2 \text{ si } (\eta_{ij}, \eta_{ij'}) = (1, 1)$$

$$= -\sigma_{\phi\theta} \text{ si } (\eta_{ij}, \eta_{ij'}) = (1, 0)$$

$$= \sigma_{\theta}^2 \text{ si } (\eta_{ij}, \eta_{ij'}) = (0, 0).$$

Par conséquent,

$$\text{V}(\hat{p}_{11}|S) = (\sum m_i^2 - n) / n^2 \text{CC} + \text{VAS} / n. \tag{A.2}$$

Finalement, en combinant (A.1) et (A.2) dans l'expression

$$\text{V}(\hat{p}_{11}) = \text{VE}(\hat{p}_{11}|S) + \text{EV}(\hat{p}_{11}|S), \text{ nous obtenons}$$

$$\text{V}(\hat{p}_{11}) = 1 / n (\text{VE} + \text{VAS}) + (\sum m_i^2 - n) / n^2 \text{CC}. \tag{A.3}$$

Si de plus, nous supposons que $m_i = n$ pour tout i , nous obtenons la forme de l'expression (11).

BIBLIOGRAPHIE

BIEMER, P. P., et STOKES, S. L. (1985). Optimal design of interviewer variance experiments in complex surveys. *Journal of American Statistical Association* 80, 158-166.

BUREAU OF THE CENSUS (1985). *Evaluating censuses of population and housing*, Statistical Training Document. United States Bureau of the Census, Washington, D.C.

BURNHAM, K. P., ANDERSON, D. R., WHITE, G. C., BROWNIE, C., et POLLOCK, K. H. (1987). *Design et Analysis Methods for Fish Survival Experiments Based on Release — Recapture*. American Fisheries Society Monograph 5.

FELLEGI, I. P., et SUNTER, A. B. (1969). A Theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

HANSEN, M. H., HURWITZ, W. N., et BERSHAD, M. A. (1961). Measurement errors in censuses and surveys. *Bulletin of the International Statistical Institute*, 38, 359-374.

REMERCIEMENTS

La rédaction du présent mémoire a été rendue possible grâce à un Joint Statistical Agree-ment avec le Bureau of the Census des Etats-Unis. L'auteur souhaite remercier Aref Dejamal du Bureau of the Census pour avoir mis à sa disposition le soutien informatique nécessaire. Il remercie également Bernice Garrett qui a dactylographié le mémoire et s'est chargée de sa relecture.

APPENDICE

Dérivation des formules de l'EQM

Soit U la population de taille N à dénombrer, U_c le sous-ensemble tiré de la population U l'échantillon tiré de l'EP, et S_c l'ensemble des personnes de l'EP faisant partie du recensement, représentant $S \cap U_c$. Enfin, soit les n éléments de l'échantillon $S u_1, \dots, u_n$. Définissons les variables

$$\eta_i = 1 \text{ si } u_i \in S_c$$

$$= 0 \text{ si } u_i \notin S_c$$

$$y_i = 1 \text{ si } u_i \text{ a été classé (par appartenance) dans } S_c.$$

$$= 0 \text{ si } u_i \text{ n'a pas été classé dans } S_c.$$

Modèle d'évaluation de l'erreur de commis

Supposons que: 1) y_i est une variable aléatoire, que les probabilités que $P(y_i = 1 | \eta_i = 0) = \phi$ et $P(y_i = 0 | \eta_i = 1) = \theta$, et 2) que y_i et y_j sont des variables indépendantes étant donné η_i et η_j pour $i \neq j$. Soit $E(\cdot | S)$ l'espérance conditionnelle et $V(\cdot | S)$, la variance conditionnelle étant donné S . Alors, $\hat{p}_{11} = \sum y_i / n$ et $E(\hat{p}_{11} | S) = (1 - \theta)\hat{p}_{11} + \phi(1 - \hat{p}_{11})$ où $\hat{p}_{11} = \sum \eta_i / n$. Le résultat de l'expression (5) est obtenu en calculant l'espérance par rapport S .

De plus, $V(y_i | \eta_i = 0) = \phi(1 - \phi)$ et $V(y_i | \eta_i = 1) = \theta(1 - \theta)$. Par conséquent, $V(\hat{p}_{11} | S) = \phi(1 - \phi)(1 - \hat{p}_{11}) / n + \theta(1 - \theta)\hat{p}_{11} / n$.

VAS (8) est obtenu en calculant l'espérance par rapport à S .

Enfin, le résultat de l'expression (6) est obtenu en combinant $VE(\hat{p}_{11} | S)$ et $EV(\hat{p}_{11} | S)$.

Modèle d'évaluation de l'erreur de commis

Soit (i, j) la $j^{\text{ième}}$ personne de la tâche du $i^{\text{ième}}$ commis et soit y_{ij} et η_{ij} définis en fonction de y_i et η_i . Supposons (1) — (3) pour le modèle d'erreur de commis. E_2 , V_2 , et C_2 respectivement, l'espérance, la variance et la covariance conditionnelles par rapport aux distributions de l'erreur de commis maintenant l'échantillon de commis fixé. Les valeurs correspondantes de l'espérance, V_1 la variance, C_1 la covariance par rapport à la sélection aléatoire des vecteurs des paramètres des commis k , comme pour l'hypothèse (3), maintenant l'échantillon S fixé. Donc,

$$E_1 E_2(\hat{p}_{11}) = E_1 \left\{ \sum_i [(1 - \theta_i) \frac{n_{11}}{n} + \phi_i \frac{n_{01}}{n}] \right\} = (1 - \theta)\hat{p}_{11} + \phi(1 - \hat{p}_{11})$$

catégorie au cours des quatre étapes sont ceux pour lesquels les données recueillies sont les plus complètes. Les tableaux 5a (cas pour lesquels les données sont complètes) et 5b (cas pour lesquels les données sont incomplètes) illustrent les résultats pondérés.

Dans le premier cas, le taux d'appariement négatif faux est de .44% tandis que le taux d'appariement positif faux est nul. Donc, aucun de ces cas n'a fait l'objet d'un appariement erroné, bien qu'un certain nombre d'entre eux aient été faussement désignés comme nonappariements. Ces données peuvent attester de la plus grande habileté du personnel responsable du rattachement à appairer les cas de l'EP. I_M est estimé à .39%, ce qui représente un indice très bas. En ce qui concerne les cas pour lesquels les données sont incomplètes, le taux d'appariement négatif faux est de .93% tandis que le taux d'appariement positif faux est de 2.18%. I_M est estimé à 1.1%, ce qui représente encore un indice assez bas. Cependant, ces données indiquent que le risque de produire de faux appariements est plus grand en ce qui concerne les cas pour lesquels l'information est incomplète.

Les données recueillies dans le cadre de cette étude indiquent que l'erreur d'appariement induit un faible biais négatif (— .4%) dans l'estimateur N , qui correspond à une sous-estimation de la valeur de N d'environ un million de personnes (en supposant que $N = 250$ millions de personnes). Le biais est seulement de — .7% même pour les cas les plus difficiles. Il serait intéressant d'analyser certains sous-groupes démographiques de la population (personnes ayant déménagé, enquêtes-subsistants et personnes habitant un appartement) afin de déterminer l'importance de l'erreur d'appariement pour ces sous-groupes.

6. RÉSUMÉ

Les modèles et les formules de l'EQM qui ont été développés dans le présent mémoire peuvent être utilisés pour évaluer l'incidence de l'erreur d'appariement sur des estimations de l'erreur d'observation du recensement. Dans le cadre du recensement des États-Unis de 1990, le biais de l'erreur d'appariement semble être l'élément le plus important de l'EQM(N). Des études préliminaires ont démontré que le biais de l'erreur d'appariement pour le recensement de 1990 est de peu d'importance, sa valeur étant inférieure à .5%. Cette estimation ne tient pas compte de l'erreur d'imputation qui s'applique à environ 1% des cas de l'EP. De plus, l'exactitude des estimations de biais repose dans une large mesure sur la véracité de l'hypothèse selon laquelle le rattachement permet d'obtenir le vrai classement d'appariement. D'autres études seront nécessaires pour déterminer la validité de cette hypothèse.

Lors de l'élaboration des formules de l'erreur quadratique moyenne totale de N , il a été supposé qu'il était peu probable que N_c comporte une erreur. Cependant, dans la pratique, il convient de soustraire de N_c une estimation du nombre d'événements faux du recensement (ou des dénombrements erronés), représentée par EB . Comme cet estimateur est obtenu à partir d'un appariement des éléments d'un échantillon de la population du recensement aux éléments de la population de l'EP, EB est aussi sujet à une erreur d'échantillonnage et à une erreur d'appariement. Par exemple, une personne peut être classée comme dénombrée de façon erronée alors qu'elle l'avait été correctement (erreur de dénombrement positif faux), ou peut être classée comme dénombrée correctement alors qu'elle l'avait été de façon erronée (erreur de dénombrement négatif faux). La méthodologie et le modèle présentés pour évaluer l'incidence des erreurs d'appariement positif faux et d'appariement négatif faux pour x_1 peuvent facilement s'appliquer à l'estimateur des dénombrements erronés. On notera que les termes des formules d'approximation de Taylor, (2) et (3), utilisées pour calculer le biais et la variance de N seront remplacés par les termes correspondant au biais et à la variance de EB . En ce qui concerne les recherches à venir, des études portant sur la variance corrélée de l'erreur d'appariement seront nécessaires pour déterminer dans quelle mesure la variance due aux commis contribue à l'erreur totale de N . Il est possible que CC^* , l'incidence maximale de l'erreur corrélée, se traduise par une surestimation considérable de l'incidence de l'erreur de commis.

Tableau 4

Resultats de l'étude de rattachement (pondérés).

Classement de rattachement		Classement du rattachement	
l'appariement initial		Appariés	
Non appariés		Non appariés	
Appariés		16 690	
Non appariés		2 178	

Tableau 5a

Resultats du rattachement pour les cas qui ont conservé leur statut jusqu'à la fin du procédé de rattachement.

Classement de rattachement		Classement du rattachement	
l'appariement initial		Appariés	
Non appariés		Non appariés	
Appariés		14 458	
Non appariés		1 775	

Tableau 5b

Resultats du rattachement pour les cas dont le classement a été modifié au cours d'au moins une des quatre étapes.

Classement de rattachement		Classement du rattachement	
l'appariement initial		Appariés	
Non appariés		Non appariés	
Appariés		2 223	
Non appariés		403	

En utilisant la méthodologie de la section précédente, il est possible d'estimer Biaisrel (β_{11}), Biaisrel (N) et I_M , l'indice d'incohérence de l'appariement. Le tableau 4 révèle les résultats de l'étude du rattachement, pondérés en fonction des probabilités de sélection de l'échantillon du rattachement. Pour ce tableau, le Biaisrel (β_{11}) est estimé à $- .4\%$, et Biaisrel (N) est donc estimé à $.4\%$. Ces calculs sont effectués à l'aide de la formule (2) en supposant que le coefficient de variation de β_{11} est de 1% et en remplaçant le Biaisrel (β_{11}) par son estimation. L'indice d'incohérence de l'appariement est estimé à $.49\%$, ce qui représente un très faible pourcentage. Le taux d'appariement positif faux, ϕ , est égal à $.004$ et le taux d'appariement négatif faux, θ , est égal à $.005$.

Comme il a été mentionné à la deuxième section, la probabilité d'erreur d'appariement peut, entre autres choses, être fonction de l'exhaustivité des données de l'EP ou du recensement. L'échantillon du rattachement a été divisé en deux sous-échantillons afin d'illustrer dans quelle mesure les taux d'erreur d'appariement peuvent varier. Le premier sous-échantillon était composé des cas qui avaient été classés comme «appariés» ou «non appariés» à toutes les étapes d'appariement sans exception. Le second sous-échantillon était constitué du reste de l'échantillon, c'est-à-dire des cas dont le classement a été modifié au cours d'au moins une des quatre étapes. Cette division de l'échantillon correspond à une division reposant sur l'exhaustivité des données d'appariement étant donné que la plupart des cas classés dans la même

Les méthodes utilisées pour attribuer les codes «apparié» ou «non apparié» aux cas non résolus sont décrites dans le rapport de Schenker (1987). Ces cas, qui représentent environ 1 % de l'échantillon de l'EP, ne sont pas pris en compte dans les tableaux qui suivent puisqu'ils ont des statuts d'appariement n'avaient pas été attribués aux cas non résolus, au moment de ce test. Néanmoins, l'erreur d'imputation peut constituer une source importante d'erreur d'appariement et poser des problèmes d'évaluation particuliers. Il est probable, par exemple, que certains cas d'EP non résolus resteront non résolus lors du rattachement et aucune estimation directe de l'erreur de classement ne peut être calculée pour ces cas. Dans le test décrit plus loin, 83 % des cas non résolus dans le cadre de l'EP sont demeurés non résolus au rattachement. Inversement, 41 % des cas qui n'ont pas été résolus au rattachement, ont été résolus à l'appariement de l'EP. Si on suppose que les imputations pour les cas qui n'ont pas été résolus lors du rattachement sont erronées, on peut obtenir un majorant de l'erreur d'imputation. De même, il est possible d'obtenir un minorant en supposant que toutes ces imputations sont justes. Cependant, à moins que la proportion représentée par les imputations soit très faible, cette analyse des «meilleurs cas, pires cas» peut produire des bornes trop élevées pour être utiles.

En 1986, un test d'essai de ces méthodes d'appariement de l'EP a été effectué à Los Angeles. Un échantillon d'environ 4 000 personnes a été apparié avec le recensement d'essai de Los Angeles. Puis apparié de nouveau par des spécialistes du recensement en vue d'évaluer le biais d'appariement. Des méthodes spéciales ont été utilisées lors du rattachement afin d'assurer un classement d'appariement très exact. Le tableau 2 illustre les taux de désaccord entre les quatre étapes d'appariement et de rattachement. Ce tableau met en lumière l'amélioration des classements à chaque étape supérieure, amélioration indiquée par un taux de désaccord décroissant dans la colonne réservée au rattachement. Les données indiquent également que peu de classements sont modifiées lors de l'étape subséquente au suivi sur place (.68 % en désaccord avec l'étape 3). De plus, le TDB de l'étape finale (par rapport au rattachement) est très bas, soit de moins de 1 %.

En supposant que le processus de rattachement conduit à l'attribution de vrais statuts d'appariement, le tableau 3 indique les estimations de θ , la probabilité d'erreur d'appariement négatif faux, et de ϕ , la probabilité d'erreur d'appariement positif faux, pour chaque étape d'appariement. Il ressort de ce tableau que le taux de non-appariement faux est plus élevé dans le cas de l'appariement par ordinateur et dans le cas du premier appariement effectué par des commis. Cependant, le contraire est vrai pour les deux dernières étapes d'appariement.

Tableau 2
Comparaison des taux de désaccord entre les étapes d'appariement (%)

	Etape 2	Etape 3	Etape 4	Rattachement
Etape 1	2.9	4.4	4.7	5.5
Etape 2	0	3.3	4.0	4.8
Etape 3	3.3	0	.68	1.6
Etape 4	4.0	.68	0	.87

Tableau 3
Estimations de θ et ϕ pour les étapes d'appariement

Etape	Estimation de θ (x100%)	Estimation de ϕ (x100%)	d'appariement (taux de non-appariement faux)
1	6.2	2.3	2.3
2	5.1	3.3	3.3
3	1.5	2.1	2.1
4	.1	.3	.3

Hypothèse 3. Le rattachement comprend moins d'erreurs mais n'est pas parfait.

Supposons que $0 < \theta_B < \theta_A$ et $0 < \phi_B < \phi_A$; c'est-à-dire que les probabilités d'erreur de classement sont plus faibles dans le cas du rattachement que dans celui de l'appariement initial, sans être nulles. Il n'existe alors pas d'estimateur non biaisé du Biais (\hat{p}_{11}). Cependant, $E(TDN)$ sera plus petite que $|\text{Biais}(\hat{p}_{11})|$ si les différences $\mu_A - p_{11}$ et $\mu_B - p_{11}$ ont toutes deux le même signe; c'est-à-dire que les estimateurs définis à partir de l'appariement initial et du rattachement sont biaisés dans la même direction. Ainsi, dans ces conditions, $|TDN|$ continue l'estimateur minime de $|\text{Biais}(\hat{p}_{11})|$.

De plus, il n'existe pas d'estimateur non biaisé de VAS. Il est toutefois possible de démontrer à partir de la formule (22) que

$$E(TDB) - 2VAS = p_{11}(\theta_B - \theta_A)(1 - 2\theta_A) + (1 - p_{11})(\phi_B - \phi_A)(1 - 2\phi_A).$$

Ainsi, chaque fois que θ_A et ϕ_A ont tous deux une valeur inférieure à .5, ce qui est vrai pour la plupart des applications pratiques, il s'ensuit que

$$E(TDB) < 2VAS$$

et que I_M défini par la formule (24) constitue une sous-estimation de I_M .

5. APPLICATION AU RECENSEMENT DE 1990

Dans le cadre du recensement de 1990, l'échantillon de l'EP sera constitué d'environ 5 000 «lots» ou groupes d'environ 30 logements contigus et on tentera de procéder à l'appariement de chaque personne comprise dans les lots avec les personnes recensées. Les variables utilisées aux fins de l'appariement seront le nom, l'adresse, le lien avec le chef de ménage, le sexe, la date de naissance, l'état matrimonial, la race et l'origine hispanique. Le processus d'appariement comportera les quatre étapes suivantes.

Etape 1. Un appariement sera réalisé par ordinateur selon la technique de Fellegi et Sunter (1969). Chaque personne faisant partie de l'EP sera classée par ordinateur soit dans la catégorie des personnes appariées à une personne recensée, soit dans celle des personnes non appariées ou encore dans la catégorie des personnes qu'il serait peut être possible d'apparier (c'est-à-dire dont le classement doit être révisé par un commis).

Etape 2. Des commis effectueront une première révision afin de corriger toutes les erreurs d'appariement ou de non-appariement erroné introduites par l'ordinateur. De plus, un ensemble normalisé de règles d'appariement sera appliqué à chaque appariement possible. Ainsi, chaque personne dénombrée dans le cadre de l'EP sera classée sous une des rubriques appariement, non-appariement, appariement possible ou cas non résolu.

Etape 3. Des commis effectueront avec discernement une deuxième révision afin de vérifier de nouveau les classements réalisés aux deux étapes précédentes. A cette étape-ci, les commis, désignés sous le nom d'équipe d'appariement spéciale (EAS), peuvent aussi décider du bien fondé d'un suivi sur place auprès de certains ménages.

Etape 4. Une révision «subséquente au suivi sur place». Les cas sont examinés de nouveau à la lumière de toute information supplémentaire obtenue dans le cadre du suivi sur place. Les codes de classement définitifs sont: apparié (dénombré), non apparié (non dénombré) ou non résolu (statut d'appariement devant être déterminé à l'étape de traitement final).

Hypothèse 1. L'appariement et le rattachement ont été réalisés dans les mêmes conditions générales.

Supposons que $\theta_A = \theta_B = \theta$ et $\phi_A = \phi_B = \phi$, c'est-à-dire que le taux d'erreur de classement prévu est le même pour les deux appariements. Alors, selon la formule (21), $E(NDR) = 0$ et aucune estimation du Biais (\hat{p}_{11}) ne peut être calculée à l'aide des données. Cependant, selon les formules (22) et (8)

$$\frac{1}{2} E(TDB) = VAS \quad (23)$$

De plus, un estimateur de I_M dans la formule (9) est obtenu par

$$I_M = TDB / [2\hat{p}_{11} (1 - \hat{p}_{11})] \quad (24)$$

où \hat{p}_{11} est l'estimateur EP de p_{11} comme il a été défini pour (2). Comme alternative, un estimateur de $E(\hat{p}_{11})$ peut être obtenu à partir du tableau 1; par exemple, se reporter aux estimateurs des formules (19) et (20).

Hypothèse 2. Rattachement parfait.

Supposons que $\theta_B = \phi_B = 0$; c'est-à-dire que le rattachement est réalisé sans erreur de classement. Alors, selon la formule (21),

$$E(TDN) = -p_{11}\theta_A + (1 - p_{11})\phi_A$$

$$= \text{Biais } (\hat{p}_{11}). \quad (25)$$

De plus, la probabilité d'erreur négative fausse, θ_A , est estimée par

$$\hat{\theta} = c / (a + c). \quad (26)$$

et la probabilité d'erreur positive fausse, ϕ_A , est estimée par

$$\hat{\phi} = b / (b + d). \quad (27)$$

Un estimateur de VAS est obtenu à l'aide de la formule

$$\widehat{VAS} = \frac{1}{r} \left(\frac{a+c}{ac} + \frac{b+d}{bd} \right) \quad (28)$$

et on obtient un estimateur de I_M grâce à la formule

$$\widehat{I}_M = \widehat{VAS} / \hat{p}_{11} (1 - \hat{p}_{11}) \quad (29)$$

où \hat{p}_{11} représente un estimateur de $E(\hat{p}_{11})$ obtenue à partir de l'EP ou du tableau 1.

Alors,

(15)
$$\mu_a = p_{11} (1 - \theta_A) (1 - \theta_B) + (1 - p_{11}) \phi_A \phi_B$$

(16)
$$\mu_b = p_{11} (1 - \theta_A) \theta_B + (1 - p_{11}) \phi_A (1 - \phi_B)$$

(17)
$$\mu_c = p_{11} \theta_A (1 - \theta_B) + (1 - p_{11}) (1 - \phi_A) \phi_B$$

(18)
$$\mu_d = p_{11} \theta_A \theta_B + (1 - p_{11}) (1 - \phi_A) (1 - \phi_B)$$

où l'indice A représente l'appariement initial et l'indice B le rappairement. Définissons,

(19)
$$\mu_a = E \left(\frac{r}{a+b} \right) = p_{11} (1 - \theta_A) + (1 - p_{11}) \phi_A$$

et

(20)
$$\mu_b = E \left(\frac{r}{a+c} \right) = p_{11} (1 - \theta_B) + (1 - p_{11}) \phi_B.$$

On notera que μ_a et μ_b sont les valeurs prévues des estimations de p_{11} réalisées respectivement à partir des classements originaux et des classements du rappairement. La différence entre ces deux estimations de p_{11} , c'est-à-dire $(b - c) / r$ est appelée *taux de différence net* (TDN). Sa valeur prévue est

(21)
$$E(NDR) = \mu_a - \mu_b = -p_{11}(\theta_A - \theta_B) + (1 - p_{11})(\phi_A - \phi_B).$$

Enfin, la proportion des personnes de l'échantillon r dont le classement du rappairement ne correspond pas au classement du premier appariement est représentée par $(b + c) / r$, qui tient désigné sous le nom de *taux de différence brut* (TDB). Sa valeur prévue est

$$E(TDB) = \mu_b + \mu_c$$

(22)
$$= p_{11} [\theta_A (1 - \theta_B) + (1 - \theta_A) \theta_B] + (1 - p_{11}) [(1 - \phi_A) \phi_B + \phi_A (1 - \phi_B)].$$

Trois jeux d'hypothèses devront être pris en considération pour l'estimation des composantes de $\text{Var}(\hat{p}_{11})$ et du Biais(\hat{p}_{11}) dans l'étude du rappairement. Dans le premier cas, on suppose que l'étude de rappairement est réalisée dans les mêmes conditions générales que l'étude d'appariement initiale, de sorte que les paramètres d'erreur associés aux deux classements sont sensiblement les mêmes. Par exemple, les commis ont reçu la même formation, possèdent les mêmes aptitudes et utilisent les mêmes méthodes. Selon la deuxième hypothèse, le rappairement est parfait, c'est-à-dire que le classement du rappairement peut être considéré comme le vrai classement. La troisième hypothèse se situe quelque part entre la première et la deuxième, elle suppose que des méthodes d'appariement améliorées et plus détaillées sont utilisées pour le rappairement; cependant, nous ne sommes pas prêts à admettre que les classements du rappairement ne comportent pas d'erreur. Nous supposons plutôt qu'il se produit moins d'erreur lors du rappairement que lors de l'appariement initial.

Données d'une étude de rattachement			
Classement		Classement de rattachement	
original	Eléments	Eléments	Eléments
	appariés	appariés	non appariés
Eléments Appariés		a	
Eléments non appariés		c	
		d	

Tableau 1

Les données (non pondérées) recueillies dans le cadre d'une étude de rattachement peuvent être présentées sous la forme illustrée au tableau 1.

Supposons que l'échantillon utilisé pour le rattachement constitue un échantillon aléatoire simple de r personnes de l'EP. De plus, il est possible de supposer que le modèle d'erreur non corrigée ou le modèle d'erreur de commis de la section précédente s'applique aussi bien à l'appariement qu'au rattachement. Soit μ_t ($t = a, b, c, d$) la proportion moyenne de la case correspondante à t dans le tableau 1.

La présente section porte surtout sur l'analyse des données relatives aux études de rattachement qui constituent la méthode d'évaluation de l'erreur d'appariement la plus utilisée. Il existe deux types d'étude de rattachement. La première tente de reproduire l'appariement initial pour un échantillon de cas en utilisant les mêmes méthodes et règles d'appariement, en donnant la même formation, etc. Ce type de rattachement a pour objet d'estimer VAS, la variance d'appariement simple ou, de manière équivalente, I_M , l'indice d'incohérence de l'appariement. Le second type de rattachement est utilisé en vue d'obtenir l'appariement le plus exact possible. En conséquence, sa réalisation nécessite la mise en oeuvre de méthodes plus détaillées de même que le recours à des commis experts hautement qualifiés et à l'arbitrage, c'est-à-dire à la résolution des problèmes de non-correspondance entre les classements originaux et ceux résultant du rattachement par un tiers, expert en matière d'appariement. Ce type de rattachement a pour objet d'estimer la valeur du biais d'appariement. De plus, comme il sera démontré plus loin, il est aussi possible d'obtenir une estimation de VAS à partir de ces données.

Plusieurs ouvrages traitent de façon exhaustive des méthodes utilisées pour estimer les composantes de l'erreur de réponse dans des enquêtes-échantillons [se reporter, par exemple, aux ouvrages de Hansen, Hurwitz, Pritzker (1964), Hansen, Hurwitz, Bershad (1961)]. Les méthodes mises en oeuvre pour estimer les composantes de l'erreur d'appariement sont essentiellement les mêmes. Par exemple, pour qu'il soit possible d'estimer la composante corrigée de la variance d'appariement, CC, les tâches des commis doivent «s'interpénétrer». Cette méthode, décrite en détail dans l'ouvrage de Kish (1962), randomise l'attribution des cas d'EP aux commis de sorte que chaque tâche de commis comporte le même nombre prévu de personnes appariées. Ensuite, un estimateur de CC est obtenu en faisant la différence entre l'erreur quadratique moyenne des commis entre eux et l'erreur quadratique moyenne intra-commis tirées de l'analyse de la variance due aux commis. Pour de plus amples renseignements concernant cette méthode, se reporter à l'ouvrage intitulé Bureau of the Census (1985).

4. ESTIMATION A PARTIR D'ETUDES DE RATTACHEMENT

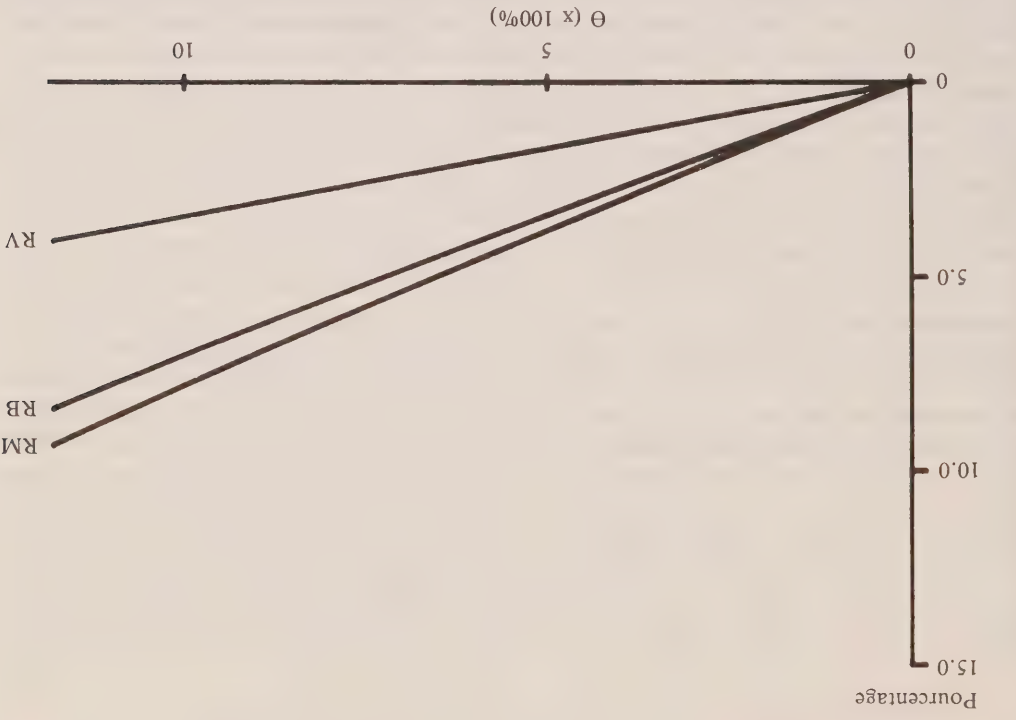


Figure 3. $RM(\Theta; \gamma)$, $RV(\Theta; \gamma)$, et $RB(\Theta; \gamma)$, comme fonctions de Θ pour $\gamma = .5$

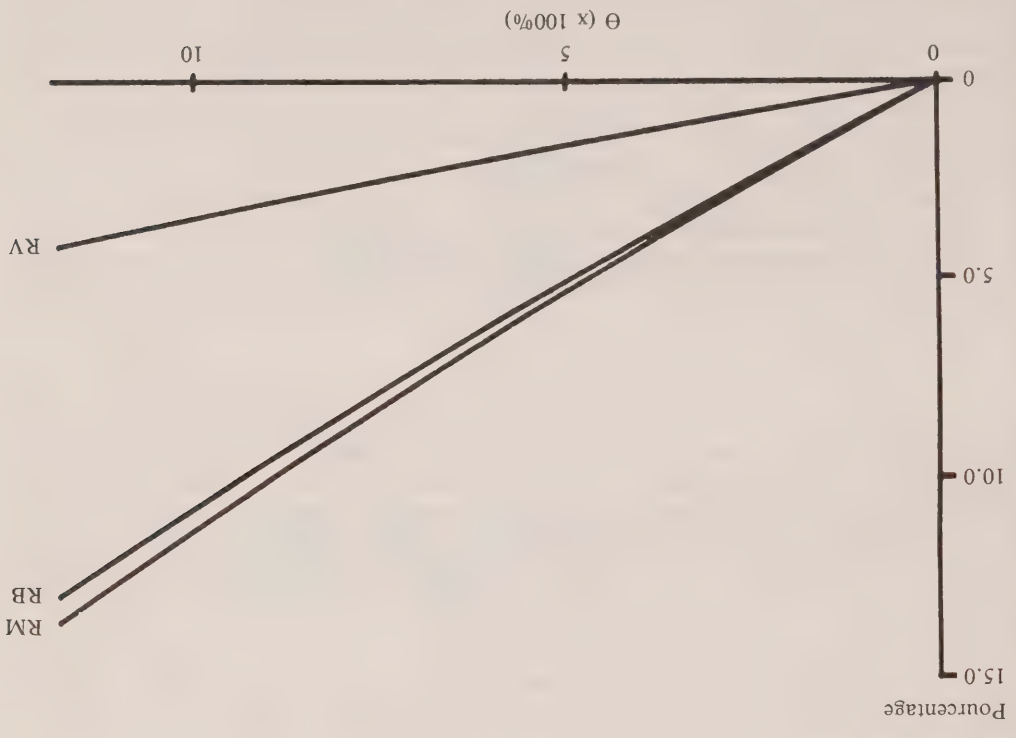


Figure 4. $RM(\Theta; \gamma)$, $RV(\Theta; \gamma)$, et $RB(\Theta; \gamma)$, comme fonctions de Θ pour $\gamma = 5$

d'erreur prévues θ et ϕ . Algébriquement, la valeur maximale de p_M est obtenue à l'aide de la formule

$$p_M^* = CC^* / (VAS + VE) \tag{14}$$

où $CC^* = p_{11}^2 \theta^2 (1 - \theta) / (1 + \theta) + (1 - p_{11})^2 \phi^2 (1 - \phi) / (1 + \phi)$ (se reporter à Johnson et Kotz (1970) pour connaître la théorie sous-jacente). Si θ_i et ϕ_i sont corrélés de façon positive, l'hypothèse admettant une corrélation égale à zéro ne fait qu'accroître l'incidence de CC. En conséquence, les exemples suivants indiquent l'incidence maximale de la variance d'appariement sur les estimations.

Afin d'illustrer l'incidence maximale d'une variance corrélée sur l'exacitude de p_{11} , on a tracé une courbe de la variation du coefficient de variation de p_{11} , soit $CV(p_{11})$, en fonction de θ pour différentes valeurs de γ . Aux fins de ces calculs, p_M^* a été substitué à p_M dans l'équation (13). L'étendue de θ était $0 \leq \theta \leq .10$ et celle de γ était $.5 \leq \gamma \leq 5$; c'est-à-dire $\phi = .20$ à $\phi = 2\theta$. Cette étendue des valeurs de γ semble acceptable puisqu, typiquement, ϕ est plus petit que θ . La figure 1 illustre la fonction lorsque $\gamma = 1$. Il n'existe pas de différence visible pour les autres valeurs de γ comprises dans l'étendue qui nous intéresse. Il semble donc que la grandeur de ϕ n'a qu'une influence négligeable sur le $CV(p_{11})$. De fait, l'expression de CC^* démontre que, lorsque $p_{11} = .85$, pas plus de 3% de la variance corrélée est attribuable à la variance de ϕ , même lorsque la valeur de ϕ est égale à celle de θ . La figure 1 laisse aussi supposer que la valeur du $CV(p_{11})$ peut passer à 2%, c'est-à-dire être doublée, pour les valeurs de θ de l'ordre de 5%.

La figure 2 illustre le biais relatif de p_{11} , représenté par $BR(p_{11})$, pour la même étendue de θ ($0 \leq \theta \leq .1$) et de γ ($.5 \leq \gamma \leq 5$). Le graphique indique clairement que le biais est moins marqué pour les valeurs plus petites de γ . De fait, le biais est égal à zéro lorsque $\gamma = (1 - p_{11}) / p_{11}$ ou à .18 en supposant, comme dans le présent exemple, que $p_{11} = .85$. Lorsque θ prend une valeur aussi faible que 5%, le biais relatif se situe entre -2% et -4%, selon la grandeur de γ . Si on compare ce résultat à l'augmentation maximale de $CV(p_{11})$ (un point de pourcentage), il appert que le biais peut avoir une incidence beaucoup plus marquée que la variance corrélée.

Afin de déterminer l'incidence possible de l'erreur d'appariement sur N , on a calculé l'augmentation de l'erreur totale en fonction de θ et pour certaines valeurs de γ . Soit $M(\theta; \gamma)$, $V(\theta; \gamma)$, et $B(\theta; \gamma)$ représentant respectivement l'erreur quadratique moyenne, la variance et le biais de N pour des valeurs données de θ et γ . $M(0; \gamma)$ représente l'erreur quadratique moyenne de N sans erreur d'appariement (c'est-à-dire $\theta = \phi = 0$) et $M(\theta; \gamma)$ représente donc de façon approximative l'erreur type de N . Définissons $RM(\theta; \gamma) = (M(\theta; \gamma) / M(0; \gamma) - 1)^{1/2}$; $RV(\theta; \gamma) = (V(\theta; \gamma) / M(0; \gamma) - 1)^{1/2}$; et $RB(\theta; \gamma) = (B^2(\theta; \gamma) / M(0; \gamma))^{1/2}$.

Ainsi, $RM(\theta; \gamma)$ est égal au rapport de la racine carrée de l'augmentation de l'erreur quadratique moyenne totale de N pour un θ et un γ donné, à la racine carrée de EQM de N sans erreur d'appariement. $RV(\theta; \gamma)$ représente la proportion de cet accroissement attribuable à la variance d'appariement, tandis que $RB(\theta; \gamma)$ représente la proportion attribuable au biais d'appariement. Par conséquent, $RM(\theta; \gamma)^2 = RV(\theta; \gamma)^2 + RB(\theta; \gamma)^2$. Les figures 3 et 4 illustrent ces fonctions pour deux valeurs extrêmes de γ , $\gamma = .5$ et 5 et pour $0 \leq \theta \leq .1$. La valeur maximale de la variance corrélée, CC^* , a une fois de plus été utilisée aux fins du calcul de la variance. Aussi, il est probable que la contribution de la variance d'appariement à l'erreur totale soit exagérée de façon considérable.

Ces figures démontrent que, pour ces valeurs de θ et de γ , la majeure partie de l'erreur est attribuable au biais, bien que la contribution de la variance puisse être non négligeable. De plus, comme il a été mentionné pour les figures 1 et 2, le biais d'appariement compte pour la majeure partie de l'erreur d'appariement totale chaque fois que le pourcentage d'erreurs négatives fausses est supérieur à celui des erreurs positives fausses.

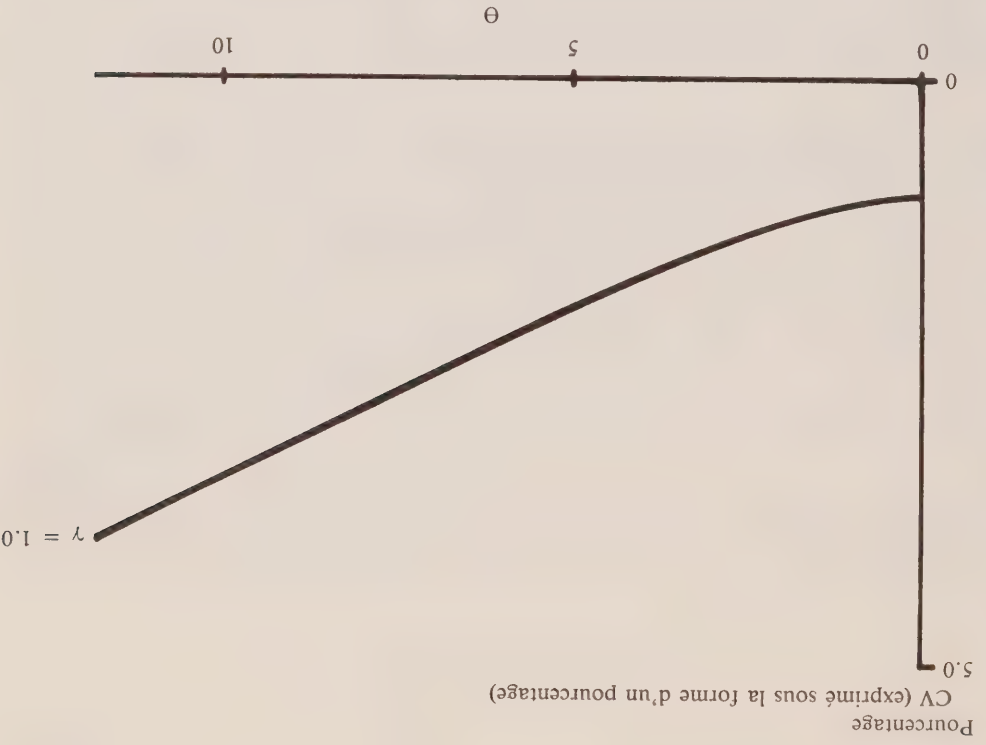


Figure 1. Coefficient de variation de p_{11} représenté comme une fonction de θ pour $\gamma = 1$

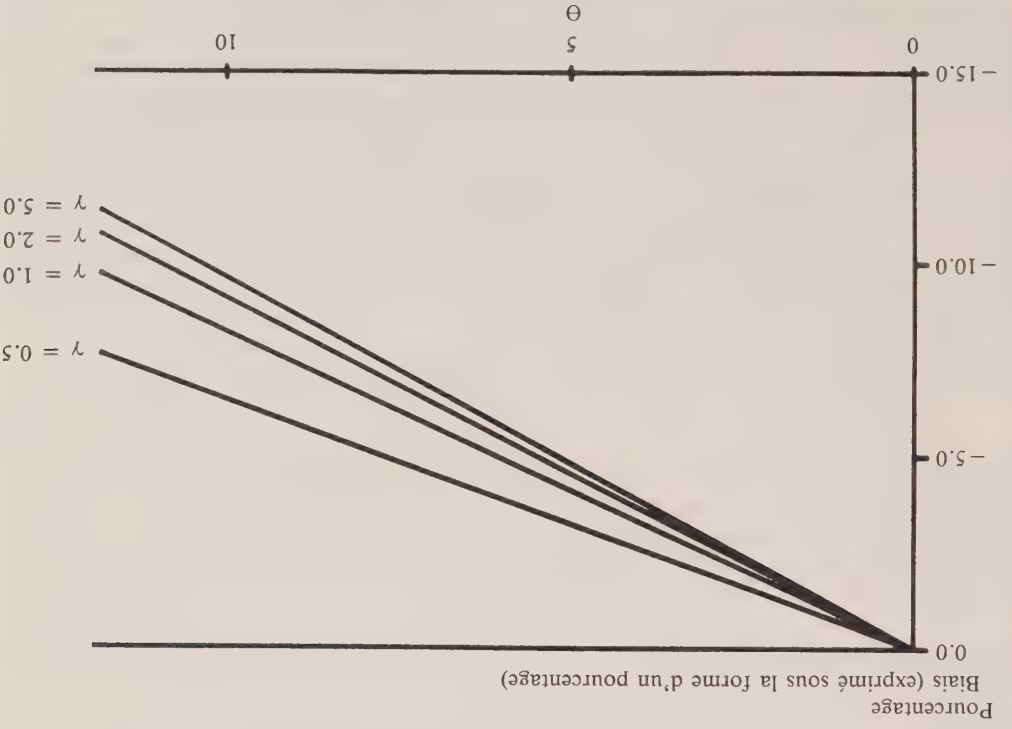


Figure 2. Biais relatif de p_{11} représenté comme une fonction de θ pour des valeurs données de γ

l'exhaustivité des renseignements d'appariement, le fait que la personne ait déménagé ou non, qu'elle habite une maison unifamiliale ou un appartement, etc.

Une façon simple de modéliser cette situation consiste à stratifier l'échantillon de l'EP en fonction d'une variable, disons Z , corrélée avec les probabilités d'erreur de classement α_j . La variable Z peut être un indicateur de l'exhaustivité des renseignements, du type d'élément, etc. Supposons que L de ces sous-populations sont identifiées par l'indice h . Soit (i, h, j) le $j^{\text{ème}}$ élément de la $h^{\text{ème}}$ sous-population dans la tâche du $i^{\text{ème}}$ commis où $i = 1, \dots, K_j$; $h = 1, \dots, L$, $j = 0, \dots, m_{ih}$; et m_{ih} représente le nombre d'éléments de la tâche du $i^{\text{ème}}$ commis compris dans la sous-population h . Comme dans le cas du modèle de l'erreur de commis, l'hypothèse (1) peut encore être retenue. Cependant, supposons en plus que:

2. $\alpha_{ihj} = \theta_{ih}$ si la personne (i, h, j) a vraiment été recensée.

$= \phi_{ih}$ si la personne (i, h, j) n'a pas vraiment été recensée.

3. $E(\theta_{ih}) = \theta_h$; $E(\phi_{ih}) = \phi_h$

$\text{Var}(\theta_{ih}) = \sigma_{\theta h}^2$; $\text{Var}(\phi_{ih}) = \sigma_{\phi h}^2$;

$\text{Cov}(\theta_{ih}, \theta_{ih'}) = \sigma_{\theta h}^2$ si $h = h'$

$= 0$ si $h \neq h'$

Selon ces hypothèses, nous obtenons $\text{Biais}(\hat{p}_{11}) = \Sigma \pi_h \text{Biais}(\hat{p}_{11})$ et $\text{Var}(\hat{p}_{11}) = \Sigma \pi_h^2 \text{Var}(\hat{p}_{11h}) + \Sigma \pi_h [E(\hat{p}_{11h}) - E(\hat{p}_{11})]^2$ où $\text{Biais}(\hat{p}_{11h})$, $E(\hat{p}_{11h})$, et $\text{Var}(\hat{p}_{11h})$ sont respectivement obtenus à l'aide des formules (5), (4), et (6), indiquant les paramètres de l'erreur de commis et p_{11} par h et où $\pi_h = E(n_h/n)$, la proportion de la population comprise dans la $h^{\text{ème}}$ sous-population.

3. DÉMONSTRATION DE L'INCIDENCE DE L'ERREUR D'APPARIEMENT SUR L'ERREUR TOTALE

Les modèles définis à la section précédente peuvent être utilisés pour démontrer l'incidence de l'erreur d'appariement sur l'erreur quadratique moyenne totale de N et p_{11} . Dans l'exemple qui suit, nous donnerons aux paramètres du modèle des valeurs hypothétiques qui, selon notre expérience, représentent des valeurs types et sont compatibles avec les paramètres de conception de l'EP de 1990.

Dans le cadre de l'EP, des estimations de N seront réalisées pour un certain nombre de sous-populations du recensement. Assumons que le coefficient de variation désiré des estimations est de 1%. Des équipes de commis effectueront l'appariement dans un certain nombre de centres de traitement. (On trouvera des renseignements plus détaillés sur le processus d'appariement dans la section suivante.) Afin d'illustrer l'effet de l'erreur d'appariement sur l'ESD, considérons une sous-population «type» de l'EP. Admettons, pour cette sous-population, que $p_{11} = .85$ et que K , le nombre de commis préposés à l'appariement dans un centre de traitement, est égal à 10. Dans notre analyse, nous avons considéré des valeurs de θ variant entre 0 et .10 ainsi qu'un certain nombre de valeurs types pour le rapport $\gamma = \theta / \phi$, c'est-à-dire, le rapport de la probabilité des erreurs d'appariement négatif faux à la probabilité des erreurs d'appariement positif faux. Étant donné que la mesure de p_M pour une erreur d'appariement n'a jamais été effectuée dans le cadre d'aucune étude, il n'existe que peu d'information susceptible d'indiquer l'étendue type de p_M . Cependant, si nous assumons que les probabilités d'erreur de commis θ_i et ϕ_i suivent une distribution Bêta unimodale et qu'elles sont corrélées, on peut obtenir une valeur maximale de p_M correspondant aux valeurs données des probabilités

Cette hypothèse est analogue aux hypothèses formulées à l'égard des erreurs d'intervieweur dans le cadre des modèles d'analyse de l'incidence des erreurs d'intervieweur [se reporter, par exemple, à Kish (1962), Hartley et Rao (1978) ainsi que Biemer et Stokes (1985)]. Il convient d'utiliser cette hypothèse si l'objectif visé est d'estimer les paramètres d'un groupe de commis beaucoup plus important dont les k commis de l'EP constituent un échantillon représentatif. Il est démontré dans l'appendice que $E(\hat{\beta}_{11})$ est quand même obtenu par l'équation (4) en admettant un sondage aléatoire simple. La formule générale utilisée pour déterminer $\text{Var}(\hat{\beta}_{11})$ correspond à la formule (A.3) de l'appendice; cependant, une simplification utile est obtenue si on peut supposer que la taille des tâches m_i est approximativement égale à m , la taille moyenne, et que le nombre d'appariements prévus est le même pour chaque tâche de commis (c'est-à-dire que les tâches de commis s'interpénètrent). Alors,

(11)

$$\text{Var}(\hat{\beta}_{11}) = \frac{1}{n} (SV + SMV) + \frac{m-1}{m} \frac{1}{k} CC$$

où CC , représentant la composante corrélée de la variance d'appariement, est

(12)

$$CC = p_{11}^2 \sigma_\theta^2 + (1 - p_{11})^2 \sigma_\phi^2 - 2p_{11}(1 - p_{11}) \sigma_{\theta\phi}$$

et VE , VAS sont obtenues respectivement à l'aide des équations (7) et (8).

Il est à noter que CC résulte de la variabilité des probabilités d'erreur de classement θ_i et ϕ_i . De plus, si on considère que CC constitue la variance de $-p_{11}\theta_i + (1 - p_{11})\phi_i$ et que les termes de cette équation sont similaires à ceux de l'équation (5), nous constatons que CC représente la variance des biais *nets* parmi les commis. Ce dernier facteur constitue la preuve que la valeur de CC doit être positive. La variance due aux commis a donc pour effet d'accroître la valeur de la variance de $\hat{\beta}_{11}$.

En se reportant encore aux ouvrages consacrés à la variance de réponse, on peut définir un paramètre p_M analogue au coefficient de corrélation intra-intervieweur, ρ , défini par Kish (1962). Aux fins du présent mémoire, le symbole p_M note la corrélation intra-commis puisqu'il représente la corrélation entre les classements d'appariement de toute paire d'éléments compris dans la même tâche de commis. Selon le modèle,

$$p_M = \frac{CC}{SV + SMV}$$

est le rapport de la composante corrélée de la variance à la variance totale associée à un seul classement. Ce rapport peut être interprété comme le degré selon lequel les commis «influencent» sur les taux d'appariement à l'intérieur de leurs tâches respectives. Ces considérations nous permettent d'établir une autre équation, équivalente à (11), pour obtenir $\text{Var}(\hat{\beta}_{11})$, soit

(13)

$$\text{Var}(\hat{\beta}_{11}) = \frac{n}{VE + VAS} [1 + (m-1)p_M]$$

2.3 Post-stratification

Le modèle d'erreur non corrélée et le modèle d'erreur de commis tiennent tous deux pour acquis (essentiellement) que le degré de difficulté afférent à la détermination de la classe exacte d'appariement est le même pour toutes les personnes dénombrées dans le cadre de l'EP (hypothèse (2) dans le cas des 2 modèles). Dans le cas du modèle d'erreur de commis, par exemple, le vecteur de la probabilité d'erreur de classement (θ_i, ϕ_i) est le même pour toutes les unités de la tâche du *i*^{ème} commis. Toutefois, dans la pratique, certaines personnes sont beaucoup plus difficiles à classer que d'autres. Ces différences sont attribuables à des facteurs comme

Il est possible que les hypothèses (1) et (2) se révèlent trop restrictives pour certaines applications. Par exemple, l'hypothèse d'indépendance (1) est infirmée lorsque l'élément B de l'EP et l'élément A du recensement sont appariés de façon erronée, ce qui entraîne le classement erroné de l'appariement correct, l'élément A du recensement et l'élément A de l'EP, comme un non-appariement. Etant donné que cette situation implique que les erreurs relatives aux éléments A et B sont corrélées de façon négative, il s'ensuit que $\text{Var}(\hat{p}_{11})$ aura une valeur plus petite que celle calculée selon la formule (6). Cependant, les erreurs corrélées n'ont aucune incidence sur $E(\hat{p}_{11})$. Un autre type d'erreur d'appariement corrélée se produit lorsque l'appariement est réalisé par des commis faisant preuve d'une tendance variable à commettre des erreurs positives fausses et des erreurs négatives fausses. Un modèle décrivant ce type d'erreurs est présenté à la section suivante.

Il est stipulé dans l'hypothèse 2 que les probabilités d'erreur de classement α_j sont les mêmes pour toute la population de l'EP. Il peut s'agir d'une autre simplification étant donné qu'il est possible de classer certaines personnes, peut-être la majorité, avec un risque d'erreur relativement petit, tandis que l'appariement des autres personnes est plus difficile à réaliser. Fondamentalement, les erreurs d'appariement sont attribuables à l'inexactitude ou à la nature incomplète des informations concernant les caractéristiques de chaque personne dans l'un ou l'autre, ou encore dans les deux systèmes. Par conséquent, si l'échantillon EP peut être post-traité selon l'exhaustivité des renseignements devant être utilisés aux fins de l'appariement, l'hypothèse peut être plausible (du moins approximativement) pour chaque sous-population. Dans un tel cas, l'erreur d'appariement totale représente une agrégation des taux d'erreur de sous-populations. Ce modèle est étudié dans la dernière sous-section du présent mémoire.

Enfin, l'hypothèse d'un sondage aléatoire simple simplifie de beaucoup l'équation de $\text{Var}(\hat{p}_{11})$. Comme les échantillons EP sont des échantillons complexes, l'hypothèse constitue une simplification qui permet néanmoins d'élaborer des formules utiles pour: a) déterminer quelles sont les composantes de l'erreur d'appariement susceptibles d'avoir l'incidence la plus marquée sur l'EQM totale de N_j ; b) concevoir des études d'évaluation de l'erreur d'appariement et affecter les ressources nécessaires à la réalisation de ces études. Dans bien des cas, l'ajustement de VE à l'aide d'une constante «d'effet de plan» permettra de compenser pour la majeure partie l'incidence de l'échantillonnage complexe sur $\text{Var}(\hat{p}_{11})$. En outre, pour l'essentiel, les formes d'échantillonnage plus complexes que le sondage aléatoire simple n'ont aucune incidence sur $E(\hat{p}_{11})$, pour autant que l'estimateur \hat{p}_{11} soit pondéré de façon appropriée. En conséquence, la forme de $B(\hat{p}_{11})$ ne dépend pas de cette hypothèse.

2.2 Modélisation des erreurs de commis

Supposons que l'appariement des personnes dénombrées dans l'EP et dans le recensement est effectué par k commis. Soient m_i le nombre de personnes appartenant à l'EP classées par le commis i , $i = 1, \dots, k$, et l'indice double (i, j) , la $j^{\text{ième}}$ personne faisant partie de la tâche du

Supposons:

1. L'événement {élément (i, j) est mal classé} et l'événement {élément (i', j') est mal classé} sont indépendants lorsque $i \neq i'$ et conditionnellement indépendants pour tout commis i , $i = i'$, $j \neq j'$, $i = i'$, $k, j, j' = 1, \dots, m_i$.
2. $\alpha_{ij} = \theta_i$ si la personne (i, j) a vraiment été recensée, et $= \phi_i$ si la personne (i, j) n'a pas vraiment été recensée.

$$3. E(\theta_i) = \theta; E(\phi_i) = \phi; \text{Var}(\phi_i) = \sigma_\phi^2; V(\phi_i) = \sigma_\phi^2; \text{et } \text{Cov}(\theta_i, \phi_i) = \sigma_{\theta\phi}.$$

En ce qui concerne le sous-ensemble de personnes faisant partie de la tâche du $i^{\text{ième}}$ commis, les hypothèses 1 et 2 sont analogues aux hypothèses 1 et 2 du modèle exposé à la section précédente. L'hypothèse 3 précise que les probabilités d'erreur d'appariement de commis sont indépendantes et qu'elles constituent des variables aléatoires distribuées de façon identique.

À partir des équations (2) et (3), on constate que l'erreur quadratique moyenne totale (EQM) de N est fonction de l'EQM totale de p_{11} . La section suivante sera consacrée à l'étude de certains modèles d'évaluation de l'incidence de l'erreur d'appariement sur p_{11} . En admettant que j ($j = 1, \dots, n$) constitue l'indice de la $j^{\text{ème}}$ personne de l'échantillon EP, α_j peut être défini comme la probabilité qu'une erreur de classement se produise pour une personne j lors du processus d'appariement et nous pouvons étudier deux hypothèses formulées à l'égard des probabilités α_j .

2. MODELES D'ERREUR D'APPARIEMENT

2.1 Erreur d'appariement non corrélée

Supposons que:

1. L'événement {élément j est mal classé} est indépendant de l'événement {élément j' est mal classé} chaque fois que $j \neq j'$.

2. $\alpha_j = \theta$ si l'élément j a vraiment été recensé, cette probabilité est désignée comme étant la probabilité d'une erreur négative fausse, et $\alpha_j = \phi$ si l'élément j n'a pas vraiment été recensé, cette probabilité est désignée comme étant la probabilité d'une erreur positive fausse.

Afin de fixer les idées, supposons que l'EP a fait l'objet d'un sondage aléatoire simple et que n est petit comparé à N , ainsi

$$\begin{aligned} E(p_{11}) &= p_{11}(1-\theta) + (1-p_{11})\phi, \\ \text{Biais}(p_{11}) &= -p_{11}\theta + (1-p_{11})\phi \\ \text{Var}(p_{11}) &= n^{-1}E(p_{11})(1-E(p_{11})) \\ &= n^{-1}(VE + VAS), \end{aligned}$$

où VE, la variance d'échantillonnage, est obtenue par

$$\begin{aligned} VE &= p_{11}(1-p_{11})(1-\theta-\phi)^2 \\ \text{et où VAS, la variance d'appariement simple, est obtenue par} \end{aligned}$$

$$VAS = p_{11}\theta(1-\theta) + (1-p_{11})\phi(1-\phi)$$

(se reporter à l'appendice pour la preuve). Les lecteurs qui connaissent le modèle d'erreur de réponse de Hansen, Hurwitz, et Pritzker (1964) reconnaîtront la correspondance entre leur variance de réponse simple et la VAS du précédent modèle. Hansen et collaborateurs ont défini une mesure I , désignée sous le nom d'"indice d'incohérence [réponse]", comme le rapport entre la variance de réponse simple et la variance totale d'une seule réponse, c'est-à-dire la proportion de la variance qui constitue une variance de réponse. Dans le cas des réponses d'enquête, I est un indicateur de la fiabilité de réponse des informations recueillies lors de l'enquête. Il est possible d'obtenir une mesure analogue pour l'erreur d'appariement afin de déterminer l'incidence du manque de fiabilité de l'appariement sur la variance de p_{11} . Cette mesure, notée I_M , est calculée à l'aide de la formule suivante:

$$I_M = \frac{VE + VAS}{VAS}$$

(9)

où $\hat{p}_{11} = x_{11}/N^p$ est un estimateur de p_{11} , la proportion réelle de la population de l'EP dénombrée dans le recensement; Biaisrel (\hat{p}_{11}) = Biais (\hat{p}_{11}) / p_{11} ; et Varrel (\hat{p}_{11}) = $\text{Var}(\hat{p}_{11}) / E^{-2}(\hat{p}_{11})$. Ces équations supposent que N_c , le nombre de personnes recensées, a une variance égale à zéro. Il s'agit bien sûr d'une simplification puisque, comme il a déjà été mentionné, une estimation des faux événements du recensement a pu être soustraite du nombre de personnes recensées afin d'obtenir N_c , et que cette correction peut être sujette à des erreurs d'échantillonnage ou à d'autres erreurs. Néanmoins, l'hypothèse retenue est valable compte tenu de l'accent mis dans le présent mémoire sur l'erreur d'appariement et son incidence sur N . Une application de cette méthodologie permettant la prise en compte de l'erreur dans l'estimateur N_c est présentée dans la dernière section du mémoire.

$$\text{Var}(N) = N^2 \text{Varrel}(\hat{p}_{11}) [1 + \text{Biaisrel}(\hat{p}_{11})]^{-2} \quad (3)$$

$$x [1 + \text{Biaisrel}(\hat{p}_{11})]^{-1}$$

$$\text{Biais}(N) = -N[\text{Biaisrel}(\hat{p}_{11}) - \text{Varrel}(\hat{p}_{11})] \quad (2)$$

Comme nous le verrons plus loin, deux types d'erreurs peuvent avoir une incidence sur N : l'erreur d'échantillonnage et l'erreur non due à l'échantillonnage. Bien que l'erreur non due à l'échantillonnage puisse provenir de plusieurs sources, la source qui nous intéresse ici est l'erreur d'appariement; c'est-à-dire l'erreur découlant d'un classement incorrect des personnes dénombrées dans le cadre de l'EP comme ayant été recensées (erreurs positives fausses) ou comme n'ayant pas été recensées (erreurs négatives fausses). En utilisant les développements en séries de Taylor, on peut dériver les formes générales des moments de N . Il est possible de démontrer que, pour les termes de l'ordre de $1/n$, où n représente la taille de l'échantillon de l'EP,

$$N = \frac{N^p N_c}{x_{11}} \quad (1)$$

mateur de système dual ou ESD) de N est de la population de l'EP. L'estimateur Sekar-Deming (désigné depuis peu sous le nom d'estimateur de N_c et du recensement, et N^p , l'estimateur non biaisé de conception de la taille x_{11} , l'estimateur non biaisé de conception du nombre total des personnes comprises dans les appariées afin de déterminer le nombre de personnes dénombrées dans les deux cas. Soient les personnes dénombrées dans l'EP et les personnes dénombrées dans le recensement sont de l'enquête postcensitaire constituent deux événements indépendants. soustrait de N_c ; et b) que le fait d'être dénombré dans le cadre du recensement et dans le cadre (par exemple, des doubles comptes, des falsifications, des personnes hors du champ de l'enquête ou non identifiables) ou que le nombre de ces événements peut être estimé avec exactitude et recensement. Supposons : a) que ni le recensement ni l'EP ne comprennent de faux événements on réalise une enquête postcensitaire (EP) dont la période de référence est identique à celle du sous le nom d'erreur d'observation du recensement), laquelle équivaut à l'estimation de N , cadre duquel N_c personnes sont dénombrées. Afin de permettre l'estimation de $N-N_c$ (désigné Considérons une population U dont la taille est de N . Un recensement est réalisé dans le et Wolter (1986).

N . En vue d'asseoir l'étude de l'erreur d'appariement sur une base simple et connue, nous adopterons le modèle original de saisie-résaisie de Sekar-Deming. On trouvera d'autres développements de la technique de Sekar-Deming dans les ouvrages de Marks, Selzer et Krotki (1974),

Modélisation de l'erreur d'appariement et son effet sur les estimations de l'erreur d'observation du recensement

PAUL P. BIEMER¹

RÉSUMÉ

L'efficacité des estimateurs de système dual du sous-dénombrement du recensement repose en grande partie sur l'hypothèse selon laquelle des personnes dénombrées lors de l'étude d'évaluation peuvent être appariées de façon précise aux mêmes personnes dénombrées lors du recensement. Or, les erreurs d'appariement et les non-appariements erronés, qui sont inévitables, diminuent l'exactitude des estimateurs. De fait, des études ont démontré que l'ampleur de l'erreur résultante peut être suffisamment importante par rapport à l'erreur d'observation du recensement pour que l'estimation devienne inutilisable. Le présent mémoire a pour objet d'exposer un modèle d'analyse de l'effet de l'erreur d'appariement sur les estimateurs du sous-dénombrement et d'illustrer son utilisation possible dans le cadre du programme d'évaluation du sous-dénombrement du recensement de 1990. L'erreur quadratique moyenne de l'estimateur de système dual est d'abord dérivée du modèle proposé et les composantes de l'EQM résultant de l'erreur d'appariement sont définies et expliquées. Nous étudierons ensuite, à la lumière du modèle, l'incidence de l'erreur d'appariement sur l'EQM de l'estimateur du sous-dénombrement du recensement. Enfin, nous illustrerons une méthodologie permettant d'utiliser le modèle pour optimiser la conception des études d'évaluation de l'erreur d'appariement et donnerons la forme des estimateurs.

MOTS CLÉS: Sous-dénombrement; estimation de système dual; saisie-resaisie; erreur non due à l'échantillonnage; erreur de traitement.

1. INTRODUCTION

Sekar et Deming (1949) ont été les premiers à suggérer l'utilisation des méthodes de saisie-resaisie aux fins de l'évaluation du recensement et de l'enregistrement des naissances et des décès. Pour permettre l'estimation de l'erreur d'observation du recensement, la méthode nécessite l'appariement des personnes dénombrées dans le cadre d'une enquête-échantillon de la population et des personnes dénombrées dans le cadre du recensement en vue de déterminer le nombre de personnes dénombrées dans les deux cas. La mise en oeuvre de la méthode de saisie-resaisie peut se buter à nombre de difficultés susceptibles d'introduire un biais considérable dans l'estimation de la taille de la population totale, N [voir, par exemple, Burnham et collaborateurs (1987) et Wolter (1986)]. Un problème qui se pose assez souvent est l'incapacité d'apparier de façon précise les personnes dénombrées dans le cadre de l'enquête-échantillon et les personnes dénombrées dans le cadre du recensement. Selzer et Adlakta (1974) ont démontré que l'erreur d'appariement peut entraîner des biais relatifs de l'ordre de 33 % et qu'elle peut être positive ou négative selon que le nombre des faux non-appariements soit supérieur à celui des faux appariements ou l'inverse (voir aussi, Scheuren et Oh, 1985). Wolter (1983) souligne que l'une des raisons justifiant la non-corrrection de données du recensement des États-Unis de 1980 était la présomption d'erreurs d'appariement dans le cadre du programme de post-dénombrement de 1980.

Le présent mémoire fournit un cadre de base pour l'évaluation de l'erreur d'appariement afférente aux études de saisie-resaisie (en particulier pour leur application à des populations humaines) et pour l'évaluation de l'incidence de ces erreurs sur l'exactitude de l'estimation de

¹ Paul P. Biemer est chef du Département of Experimental Statistics et directeur du University Statistics Center de l'Université du Nouveau-Mexique, Las Cruces, Nouveau-Mexique, États-Unis.

BIBLIOGRAPHIE

ANOLIK, J. (1988). The rural post-enumeration survey in east central Mississippi. Statistical Research Division Report, Series RR 88/10. U.S. Bureau of the Census, Washington, D.C.

CHOI, C.Y., STEEL, D.G., et SKINNER, T.J. (1988). Adjusting the 1986 Australian Census for under-enumeration. *Proceedings of the Census Bureau Fourth Annual Research Conference*. Bureau of the Census, Washington, D.C.

CITRO, C.F., et COHEN, M.L. (1985). *The Decennial Census: New Directions for Methodology in 1990*. Washington: National Academy Press.

COWAN, C.D., et MALEC, D. (1986). Capture-recapture models when both sources have clustered observations. *Journal of the American Statistical Association*, 81, 347-353.

DIFFENDAL, G. (1988). Test des opérations de redressement de 1986 dans le Central Los Angeles County. *Techniques d'enquête*, 14.

ERIKSEN, E.P., et KADANE, J. (1985). Estimating the population in a census year: 1980 and beyond. *Journal of the American Statistical Association*, 80, 98-114.

FAY, R.E., PASSEL, J.S., ROBINSON, G., et COWAN, C.D. (1988). The coverage of population in the 1980 Census. Technical report PHC 80-E4. Bureau of the Census, Washington, D.C.

HAINER, P., HINES, C., MARTIN, E., et SHAPIRO, G. (1988). Research on improving coverage in household surveys. *Proceedings of the Fourth Annual Research Conference*, Bureau of the Census, Washington, D.C.

HINES, C. (1988). The role of participant observation research in understanding the census undercount. Paper presented at the Population Association of America Annual Meetings, New Orleans, La.

JARO, M. (1988). Advances in record linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association* (à paraître).

MARK, R.W., et SAALFELD, A.J. (1988). Programs for assuring map quality at the Bureau of the Census. *Proceedings of the Bureau of the Census Fourth Annual Research Conference*, Bureau of the Census, Washington, D.C.

MULRY, M., et SPENCER, B. (1988). Total error in dual system estimates of population size. *Proceedings of the Fourth Annual Research Conference*, Bureau of the Census, Washington, D.C.

SEKAR, C.C., et DEMING, W.E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, 44, 101-115.

SCHENKER, N. (1988). Traitement des données manquantes dans l'estimation de la couverture: le test des opérations de redressement de 1986. *Techniques d'enquête*, 14.

THOMPSON, J.H., WHITFORD, D., et STODT, D. (1987). Note de service à Howard Hogan, Sujet: Review of 1986 PES Matching, avril 21, 1987.

U.S. BUREAU OF THE CENSUS (1978). *The Current Population Survey: Design and Methodology*. Technical Paper No. 40, Washington, D.C.

U.S. DEPARTMENT OF COMMERCE (1987). Communiqué de presse, Déclaration par Undersecretary Robert Ortner, octobre 30, 1987.

WOLTER, K.M. (1986a). Some coverage error models for census data, *Journal of the American Statistical Association*, 81, 338-346.

WOLTER, K.M. (1986b). A combined coverage error model for individuals and housing units. Statistical Research Division Report Series RR 86/27, U.S. Bureau of the Census, Washington, D.C.

WOLTER, K.M. (1986c). Capture-recapture estimation in the presence of a known sex ratio. Statistical Research Division Report Series RR 86/20, U.S. Bureau of the Census, Washington, D.C.

initiale de l'EP visait entre autres à recueillir les données qui devaient permettre de dire si une personne de l'échantillon P avait été manquée lors du recensement. Nous travaillions actuellement à rendre le questionnaire plus révélateur en y incorporant d'autres questions de sélection qui visent à mieux cerner les personnes ayant déménagé. Les futures enquêtes post-censitaires comporteront également des interviews de suivi pour la plupart des personnes ayant déménagé et pour les ménages de l'échantillon P qui ont déclaré ne pas avoir déménagé mais pour qui, selon certains indices, ce ne serait pas le cas. Nous croyons que ces interviews de suivi seront un moyen de réduire au minimum le nombre de déclarations fautives.

Il faut continuer d'améliorer et de tester les procédures de contrôle qualitatif qui visent à déceler les cas de fabrication dans l'EP et à en corriger les effets. Le contrôle qualitatif devrait permettre de vérifier non seulement les noms figurant sur la liste de l'EP, mais aussi d'autres éléments d'information. Par ce contrôle, nous devons pouvoir déceler les enregistrements que l'on a fabriqués partiellement en lisant le nom sur la boîte aux lettres ou en l'obtenant du propriétaire pour ensuite fabriquer les caractéristiques. Nous sommes aussi à modifier les formulaires de suivi de l'EP pour faciliter la détection de personnes fictives.

Dans les futures enquêtes post-censitaires, nous chercherons spécialement à réduire au minimum le nombre de données manquantes en tâchant particulièrement d'éliminer la nécessité d'un suivi. Par ailleurs, plus il y aura de cas qui devront faire l'objet d'un suivi, plus grande sera la proportion de cas qui n'auront pu être résolus. Il est donc nécessaire de trouver une façon de traiter convenablement ces cas.

Malgré les résultats acceptables qu'a produit le TOR, il convient d'examiner tous les aspects de la question avant de conclure que les résultats de l'EP de 1990 seront plus près de la réalité que les résultats du recensement de cette même année. Le niveau de sous-dénombrement net observé dans le test de Los Angeles était élevé par rapport à ce que l'on devrait observer dans un recensement à l'échelle du pays. Une EP nationale aura-t-elle un degré d'erreur suffisamment faible pour produire des estimations démographiques plus précises?

Nous pensons que le recensement de 1990 révélera des secteurs qui auront un fort taux de sous-dénombrement ou, peut-être, de sur-dénombrement malgré un faible taux de sous-dénombrement net à l'échelle nationale. L'EP devrait donc produire des estimations démographiques plus justes pour les secteurs où la population est la plus difficile à dénombrer. Si, d'ici deux ans, nous perfectionnons la nouvelle version de l'EP, nous pourrions peut-être aussi obtenir des estimations démographiques plus justes pour les secteurs où la population est moins difficile à recenser.

Nous pensons aussi que l'EP comportera moins d'erreurs si le taux de sous-dénombrement est moins élevé. Les secteurs stables, caractérisés par des cartes de bonne qualité, des adresses clairement définies, un faible nombre de personnes ayant déménagé et des répondants coopératifs, causeront relativement peu de problèmes tant pour le recensement que l'EP. Les erreurs de traitement résiduelles pourraient créer un seuil de précision au-delà duquel l'EP ne pourrait aller, peu importe le taux de sous-dénombrement net réel. Nous ne pourrions vérifier cela qu'à l'exécution de l'EP de 1990. Il pourrait alors y avoir des secteurs pour lesquels les estimations de l'EP seraient plus justes que les estimations du recensement et d'autres (la plupart de ceux qui restent) pour lesquels elles seraient aussi justes sinon presque aussi justes que les estimations du recensement. La théorie statistique devrait nous permettre de trouver un moyen de produire une meilleure estimation en combinant les résultats du recensement avec ceux de l'EP.

REMERCIEMENTS

Cet article est le fruit du travail de nombreuses personnes à part les auteurs; ceux-ci tiennent à remercier Dan Childers, Carol Corby, Gregg Diffendal, Charisse Jeffries, Arona Pitter, Nathaniel Schenker, Maria Urrutia et Kirsten West. Les auteurs veulent aussi exprimer toute leur gratitude aux trois arbitres pour leurs nombreux commentaires utiles.

Tableau 8

Estimations duales pour le recensement d'essai de 1986 à Los Angeles

EP			
Dénombrés		Manqués	
343,667	298,204	38,503	5,870
44,373			
Total		51,333	388,040
Total			
Dénombrés		336,707	388,040
Manqués		51,333	388,040
Total		388,040	388,040

^a Enregistrements corrects du recensement = total des enregistrements du recensement - substitutions - enregistrements erronés.

Compte tenu des données disponibles, nous n'avons pas de moyen précis pour évaluer l'importance du biais de corrélation contenu dans les données du TOR. Néanmoins, si nous nous fondons sur les travaux évoqués ci-dessus, nous avançons que le taux de sous-dénombrement déterminé par le TOR devrait être d'au moins 2.3% de plus.

9. ERREUR ALÉATOIRE

L'erreur d'échantillonnage influe sur les estimations du nombre de cas apparés, du nombre d'enregistrements erronés et des totaux de l'échantillon P. Les chiffres du recensement et le nombre de substitutions de personnes dans le recensement reposent sur un dénombrement complet de la population (100%) et ne sont donc pas touchés par l'erreur d'échantillonnage. L'écart type estimé pour le taux de sous-dénombrement est 0.007. Ainsi, un intervalle de confiance de 95% pour le taux de sous-dénombrement est $0.09 \pm 2(0.007) = (0.076, 0.104)$ si l'on suppose une distribution normale. Diffendal (1988) présente les erreurs types estimées pour les facteurs de redressement du TOR définis par $Y = N^{++}/CEN$ et utilise un modèle axé sur les composantes de la variance pour lisser les valeurs de Y dans le but d'atténuer les effets de l'erreur d'échantillonnage. Dans la plupart des cas, le lissage a réduit de façon substantielle les erreurs types estimées, particulièrement en ce qui a trait aux domaines (ou petites régions). Nous croyons que cette forme de lissage pourra être utilisée de façon profitable dans les futures EP.

10. CONCLUSION

Après le recensement de 1980, le U.S. Bureau of the Census a réexaminé son programme d'évaluation du taux de couverture et en a relevé les lacunes. Nous avons alors mis sur pied un programme de recherches puis avons élaboré une nouvelle méthode d'évaluation du taux de couverture destinée à réduire ces lacunes. Au cours des huit dernières années, le nouveau plan de sondage de l'EP a fait l'objet de tests importants qui ont démontré sa grande supériorité par rapport au plan de l'EP de 1980. Cet article nous a donné l'occasion d'analyser les résultats de notre programme de recherches dans la perspective du TOR de 1986. On ne pourra peut-être jamais avoir une EP parfaite. Cependant, le nouveau plan de sondage ne renferme pas de lacune assez grande pour réfuter les résultats de l'EP. Pour les raisons énoncées dans la section I, nous croyons que l'effet combiné des erreurs dans l'estimation du taux de couverture dans le TOR est moindre que l'erreur contenue dans le recensement à Los Angeles. Un des principaux avantages du TOR est d'attirer notre attention sur de nouveaux sujets et des problèmes mineurs qui justifient une recherche plus poussée. Par exemple, l'interview

Dans le nouveau plan de sondage de l'EP, nous tentons de réduire au minimum le biais dû à la dépendance des enquêtes en fixant la date du début des opérations de l'EP après la date de la fin des principales opérations sur le terrain pour le recensement. Contrairement à ce qui s'est fait lors de l'EP d'avril 1980, cette façon de procéder favorisera le plus possible l'indépendance entre le recensement et l'EP. En outre, nous appliquons désormais dans nos bureaux régionaux des procédures qui favorisent une telle indépendance; par exemple, pour l'EP les intervieweurs sont affectés à une région différente de celle où ils ont travaillé (si c'est le cas) durant le recensement.

Il sera difficile de supprimer le biais de corrélation dû à l'hétérogénéité dans les futures EP. Les seules solutions qui s'offrent à nous sont une post-stratification plus efficace et une combinaison quelconque des données de l'EP et de l'AD, peut-être sur la base des quotients hommes/femmes établis par l'AD. Voir Wolter (1986c) et Choi, Steel et Skinner (1988). Au cours de la dernière décennie, nous avons expérimenté diverses méthodes de post-stratification, dont celles faisant intervenir des variables comme le mode d'occupation, le pourcentage de questionnaires du recensement retournés par la poste et l'état matrimonial. Ces méthodes présentent un certain intérêt pour l'avenir. Voir Diffendal (1988).

Les écarts de taux de sous-dénombrement observés dans le TOR sont conformes aux prévisions. Dans le recensement de la population des E.-U., les hommes ont habituellement un taux de couverture plus faible que celui des femmes. C'est ce que les résultats de l'analyse démographique indiquent constamment. Les quotients de masculinité (nombre d'hommes pour 100 femmes) observés dans le TOR sont supérieurs à ceux observés dans le recensement pour les personnes d'origine hispanique et celles qui ne sont ni d'origine hispanique ni d'origine asiatique. De même, pour le groupe de personnes de 30 à 44 ans, les quotients de masculinité sont beaucoup plus élevés dans le TOR que dans le recensement (de 1,1 à 3,4 hommes de plus pour 100 femmes). Ces résultats sont conformes à ceux de l'analyse démographique de 1980 pour l'ensemble du pays. Par conséquent, nous sommes d'avis que les quotients de masculinité observés dans le TOR sont ceux qui se rapprochent le plus des quotients réels et malgré que le gain soit limité par le biais de corrélation, l'EP permet encore d'évaluer l'écart des taux de sous-dénombrement.

Le tableau 8 donne sous forme de table de contingence les résultats du TOR de 1986 sans post-stratification. Le nombre estimé de personnes qui n'ont été recensées par aucun des deux systèmes,

$$N_{22} = 5,870,$$

est du même ordre de grandeur que le nombre de substitutions dans le recensement (5,259) et le nombre d'enregistrements erronés (6,426). Environ le huitième du nombre estimé de personnes manquées dans le recensement ($N_{12} + N_{22} = 44,373$) est rattaché à la case (2,2). Par conséquent, le chiffre du sous-dénombrement peut être attribué en majeure partie à l'estimation pure et simple et non au modèle dual.

Afin d'illustrer l'effet du biais de corrélation, doublons l'effetif de la case (2,2). Cela a pour effet de hausser le taux de sous-dénombrement estimé d'environ 1,4%. À la suite d'une analyse de l'EP de 1980, Ericksen et Kadane (1985) proposent de multiplier l'effetif de la case (2,2) par un facteur de 2,7, ce qui a pour effet de hausser le taux de sous-dénombrement estimé de 2,3%. D'autres renseignements nous éclairaient sur la question du biais de corrélation. En effet, trois anthropologistes ont prêté leur concours au U.S. Bureau of the Census pour agir à titre de participants ou d'observateurs dans le test de Los Angeles. Leurs observations ne permettent pas de mesurer directement le biais de corrélation mais elles nous révèlent jusqu'à quel point les catégories de personnes manquées dans les recensements et les EP se ressemblent. Selon le rapport des anthropologistes, des personnes pour lesquelles la probabilité de saisie est très faible tendraient à être manquées dans le recensement comme dans l'EP et par conséquent, les données du TOR pourraient être entachées d'un biais par défaut appréciable. Voir Hainer et coll. (1988) et Hines (1988).

Tableau 7

Taux de sous-dénombrement (%) pour les personnes de race noire et pour l'ensemble de la population dans les recensements de 1950, 1960 et 1980 aux E.-U., et écarts entre ces taux

Source	Personnes de race noire	Population totale	Ecart
1950			
EP	3.2	1.4	1.8
AD	9.6	4.4	5.2
1960			
EP	3.8	1.9	1.9
AD	8.3	3.3	5.0
1980			
EPa			
Bas	1.1	-1.0	2.1
Moyen	6.9	1.4	5.5
Elevé	5.7	2.1	3.6
AD	5.9	1.4	4.5

La TEP de 1980 a produit 12 séries d'estimations. Les trois estimations présentées ici pour chaque catégorie sont tirées respectivement de la série la plus faible, de la série moyenne et de la série la plus élevée par rapport au taux de sous-dénombrement total estimé.

et que N_{11} , N_{12} et N_{21} sont des estimateurs directement basés sur le plan de sondage, le biais qui découle du non-respect des hypothèses d'indépendance se retrouve uniquement dans N_{22} , l'estimateur de N_{22} .

Nous pouvons analyser le biais de corrélation présent dans le recensement de 1980 et des recensements antérieurs en comparant $N^{+}_{i,t}$ à des estimations indépendantes de la population totale tirées d'une analyse démographique (AD). Le tableau 7 contient les données pertinentes des recensements récents. Si l'on considère les estimations de l'analyse démographique comme valeurs de référence, la comparaison expose le biais total dont est entaché l'estimateur dual, notamment le biais de corrélation et les autres sources d'erreur. Nous croyons que le biais par défaut exposé dans ces estimations est largement imputable au biais de sous-échantillonnage. L'EP de 1950 a sous-estimé sérieusement la taille de la population, le taux de sous-dénombrement et l'écart entre le taux de sous-dénombrement pour la population totale et celui pour les personnes de race noire à cause probablement du biais d'hétérogénéité et du biais découlant de la dépendance des enquêtes. Il convient toutefois de souligner que si les données de l'EP de 1950 avaient servi à redresser les chiffres du recensement de la même année, l'écart entre les taux de sous-dénombrement des deux groupes serait passé de 6,4% à environ 4,6%. L'EP de 1960 a sous-estimé tout aussi sérieusement la taille de la population, le taux de sous-dénombrement et l'écart entre les taux de sous-dénombrement des deux groupes à cause probablement du biais de corrélation. Si les données de l'EP de 1960 avaient servi à redresser les chiffres du recensement de cette même année, l'écart entre les taux de sous-dénombrement serait passé de 5,3% à environ 3,4%.

Il n'y a pas eu d'enquête post-censitaire en 1970. Celle de 1980 a produit 12 séries d'estimations du taux de sous-dénombrement fondées sur les résultats des sondages d'avril et d'août et sur diverses séries d'hypothèses. Les taux de sous-dénombrement calculés au moyen de l'AD se situent à peu près au centre des séries d'estimations produites par l'EP. Le biais de corrélation n'est pas aussi évident ici que dans les EP de 1950 ou de 1960 à cause principalement des améliorations qui ont été apportées en 1980 dans le but de réduire la dépendance positive des enquêtes. Nous croyons qu'il existe encore un biais d'hétérogénéité mais que celui-ci est masqué par d'autres erreurs de l'EP et un biais imputable à une dépendance négative des enquêtes.

géographie du recensement. Le secteur de recherche devait être limité aux environs immédiats de l'adresse de la CPS mais comme les adresses de cette enquête étaient fondées sur la géographie de 1980 et la recherche se trouvait être effectuée dans un plus grand secteur. Comme le secteur de recherche pour l'échantillon P s'était élargi, celui pour l'échantillon D devait aussi s'élargir. Nous croyons que cela a engendré des incohérences entre les secteurs de recherche respectifs et, par conséquent, introduit un biais dans l'ED.

Dans le TOR, nous avons exécuté le double appariement entre les personnes des échantillons P et D à l'intérieur des blocs choisis. Tout le long de l'appariement informatisé et de l'appariement manuel, la géographie et les secteurs de recherche étaient uniformes, bien circonscrits et parfaitement sous contrôle. C'est pourquoi le problème d'équilibrage n'a pas introduit de biais notable dans les résultats du test de Los Angeles.

8. BIAIS DE CORRELATION

Pour que l'estimateur dual soit un estimateur convergent de la taille réelle de la population N^{++} , nous devons poser deux hypothèses d'indépendance:

- i) indépendance des enquêtes;
- ii) indépendance des probabilités de sélection vis-à-vis des caractéristiques des personnes.

De plus, on suppose souvent l'indépendance des probabilités de sélection des personnes mais si cette hypothèse n'est pas respectée, cela n'a pas pour effet de fausser sensiblement l'estimation de la population totale. (Wolter 1986b et Cowan et Malec 1986).

L'hypothèse de l'indépendance des enquêtes ne tient plus lorsque la fréquence de saisie d'une personne pour le recensement influe sur la probabilité de saisie de cette personne dans l'EP. L'estimateur N^{++} souffre d'un biais par défaut lorsque la probabilité de saisie dans l'EP augmente du fait que la personne a été saisie dans le recensement, et d'un biais par excès lorsque la probabilité de saisie dans l'EP diminue pour la même raison.

Les données de l'EP d'avril 1980 renferment peut-être un biais appréciable du fait que l'hypothèse de l'indépendance des enquêtes n'a pas été respectée. Cela est dû au fait que les répondants pourraient avoir confondu la CPS de mars ou d'avril avec le recensement.

L'hypothèse de l'indépendance des probabilités de sélection vis-à-vis des caractéristiques des personnes ne tient plus lorsque les probabilités de saisie dans le recensement varient d'une personne à l'autre. Le biais qui en découle (appelé biais d'hétérogénéité ou biais de corrélation) est habituellement considéré comme un biais par défaut puisque les personnes qui ont une forte probabilité de saisie dans le recensement tendent aussi à avoir une forte probabilité de saisie dans l'EP et inversement, les personnes qui ont une faible probabilité de saisie dans le recensement ont aussi tendance à avoir une faible probabilité de saisie dans l'EP.

Sekar et Deming (1949) ont proposé la post-stratification comme moyen de réduire le biais d'hétérogénéité. Dans la pratique, toutefois, il est peu probable que cette méthode s'avère pratiquement efficace; il subsiste inévitablement des différences de probabilités de saisie à l'intérieur des post-strates.

Dans le modèle dual, on estime le nombre de personnes qui ont été manquées par les deux systèmes (N_{22}) au moyen de l'expression suivante:

$$N_{22} = N_{12}N_{21}/N_{11},$$

qui correspond à l'équation (2) dans Diffendal (1988). Comme l'estimateur dual peut être exprimé sous la forme

$$N^{++} = N_{11} + N_{12} + N_{21} + N_{22},$$

devrait être porté à environ

$$\frac{411}{411 + 19,334} = .021.$$

Par conséquent, le taux de sous-dénombrement calculé initialement dans le TOR devrait être abaissé d'environ 0,5 point de pourcentage, ce qui donnerait un taux de sous-dénombrement estimé d'environ 8,5%.

7. CONCILIATION DES TAUX DE SUR-DÉNOMBREMENT ET DE SOUS-DÉNOMBREMENT BRUTS

Pour estimer l'erreur de couverture nette, les méthodes et principes servant à mesurer le taux de sur-dénombrement brut doivent être compatibles avec les méthodes et principes utilisés pour mesurer le taux de sous-dénombrement brut. C'est ce que nous appelons la condition d'"équilibre". Nous décrivons ci-dessous en termes élémentaires comment réaliser l'équilibre dans l'EP. Une façon d'aborder le problème est d'envisager l'estimateur dual sous la forme

$$N_{++} = (N_{1+}N_{+1})/N_{11},$$

où

$$N_{11} = M,$$

le nombre pondéré de personnes de l'échantillon P qui ont pu être appariées, et

$$N_{+1} = N_p,$$

le nombre pondéré de personnes comprises dans l'échantillon P. On trouvera la définition de tous les symboles dans Diffendal (1988).

Comme nous ne pouvons scruter tous les questionnaires du recensement, le nombre observé de personnes appariées (M) sera inférieur au nombre réel de concordances dans les deux systèmes. Pour maintenir les coûts à un niveau acceptable, nous limitons l'appariement pour un cas donné à un "secteur de recherche", constitué habituellement du bloc échantillonné et d'un ou deux anneaux de blocs situés autour de celui-ci.

En conséquence, le terme N_{11} est l'estimateur de kN_{*11} , où $0 \leq k \leq 1$ est la probabilité conditionnelle qu'une personne recensée soit comptée dans le bon secteur de recherche, et N_{*11} est l'estimateur que l'on obtiendrait pour N_{11} dans l'EP si on pouvait exécuter la recherche dans l'ensemble de la population.

Pour construire un estimateur convergent de la taille de la population, nous devons réduire le nombre de personnes recensées par le facteur k . Comme la recherche d'enregistrements erronés pour l'échantillon D (par ex., enregistrements répétés) se fait dans le secteur de recherche et non dans l'ensemble de la population, le terme N_{1+} est l'estimateur de kN_{*1+} , où N_{*1+} est l'estimateur que l'on obtiendrait pour N_{1+} si la recherche d'enregistrements erronés pouvait se faire dans l'ensemble de la population.

Si nous supposons que les secteurs de recherche sont uniformes, l'ED devient un estimateur convergent de N_{++} . Signalons que dans ce modèle d'équilibrage, nous n'estimons pas la probabilité k ; nous comptons plutôt sur des secteurs de recherche uniformes pour retrancher cette probabilité de l'ED.

Dans l'EP de 1980, il était impossible d'équilibrer les échantillons P et D parce qu'ils ne se chevauchaient pas. Les adresses de la CPS (ou de l'échantillon P) étaient codées suivant la

Tableau 6

Résultats de l'étude sur le nouvel appartement: échantillon D (pondéré)^a

Résultats du nouvel appartement				
Résultats initiaux				
Enregistrement	Enregistrement	Enregistrement	Cas non résolu	Total
correct	erroné			
Enregistrement correct	19,153	28	88	19,269
Enregistrement erroné	41	283	1	325
Cas non résolu	140	100	223	463
Total	19,334	411	312	20,057

^a Les données sont pondérées aux totaux de l'échantillon D.

les enregistrements répétés durant l'exécution d'autres opérations de l'EP et par conséquent, plusieurs ont dû échapper à notre attention. Dans la prochaine EP, une activité spéciale sera prévue pour déceler les enregistrements répétés.

Il arrive que le recensement fournisse si peu de renseignements sur une personne que même si celle-ci subissait une interview en bonne et due forme dans l'échantillon P, il ne serait pas possible de l'appartier à un enregistrement de l'échantillon D. Pour pallier à cette difficulté, il conviendrait de classer ces cas dans les EF de manière à obtenir une estimation juste de la population totale. Ce problème se compare au problème de l'équilibrage géographique analysé dans la section 7. Dans l'EP de 1980, il était très difficile d'être conséquent à cause du non-chevauchement des échantillons D et P; des cas similaires ont été classés comme "non appartable" dans l'échantillon D et comme "appartable" dans l'échantillon P, ce qui a introduit un biais dans l'estimateur dual. Comme les échantillons P et D se chevauchent dans le nouveau plan de sondage de l'EP, nous voyons à ce que des règles identiques s'appliquent dans les deux cas, ce qui élimine le biais.

Lors d'une autre évaluation du TOR, ainsi que dans le cadre de l'étude sur le nouvel appartement dont il a été question plus haut (voir section 2), des spécialistes du bureau chef ont réexaminé les cas de l'échantillon D contenus dans un sous-échantillon de 35 blocs. Comme dans la section 2, nous avons effectué le nouvel appartement indépendamment de l'appartement initial et nous sommes ensuite prononcé sur les cas où il y avait des écarts entre les résultats des deux opérations. En définitive, nous croyons que les résultats du nouvel appartement reflètent le plus fidèlement possible les codes de dénombrement réels des personnes de l'échantillon D tandis que les écarts entre les résultats de l'appartement initial et ceux du nouvel appartement peuvent être considérés comme une mesure du biais attribuable à l'erreur dans l'opération initiale.

Les résultats pertinents figurent dans le tableau 6. Il convient de souligner que la plupart des changements ont porté sur des cas qui avaient été classés initialement parmi les cas "non résolus". Bon nombre de ces cas sont ceux que nous avons analysés plus haut, c'est-à-dire ceux qu'il fallait classer soit parmi les cas "fictifs" ou soit parmi les cas de "non-réponse". Les données du tableau 6 nous portent à conclure que de meilleures procédures s'imposent pour identifier les cas de l'échantillon D comme fictifs. Nous travaillons actuellement à l'élaboration de nouvelles procédures qui seront mises en application dans la prochaine EP du Bureau of the Census, qui sera combinée avec une répétition générale du recensement de 1990 prévue pour 1988.

D'après l'étude sur le nouvel appartement, nous croyons que le taux initial d'EF,

$$\frac{325}{325 + 19,269} = .016$$

à l'enregistrement correspondant de l'échantillon D. Les cas non apparés de l'échantillon D doivent faire l'objet d'une interview de rappel qui n'a lieu que six mois après le jour du recensement. Le resserrement des délais d'exécution a pour effet de réduire les taux de données manquantes, d'atténuer le rôle des interviews par procuration et d'améliorer la qualité des données recueillies.

On relève quatre grandes sources d'erreur dans l'estimation du nombre d'enregistrements erronés:

- i) erreurs de réponse dans l'interview de l'échantillon D (il s'agit de l'interview de l'échantillon F pour la plupart des cas et de l'interview de rappel pour tous les autres cas), ou erreurs de codage commises par les préposés au traitement des questionnaires;
- ii) erreur commise par un intervieweur ou un membre du personnel lors de l'attribution du code géographique du bureau à une personne de l'échantillon D;
- iii) erreur dans la recherche d'enregistrements répétés;
- iv) erreurs commises en classant une personne de l'échantillon D parmi les cas qui n'offrent pas suffisamment de données pour l'appariement.

Il y a aussi les erreurs dues à la non-réponse dans l'interview de l'échantillon D, comme nous l'avons vu dans la Section 5, et l'erreur d'échantillonnage, que nous verrons dans la Section 9. Les erreurs de réponse sont souvent liées à l'identification d'un membre de l'échantillon D comme une personne "fictive". Parfois, l'intervieweur constate que la personne qui demeure dans une unité de logement (ou un autre répondant admissible) ne connaît pas les personnes dont le nom figure sur la liste du recensement pour cette adresse. En règle générale, le répondant en question est une personne qui est déménagée à cette adresse après le recensement et qui ne connaît tout simplement pas les personnes qui occupaient le logement le jour du recensement. Ces cas de l'échantillon D devraient être identifiés comme des cas de non-réponse. Toutefois, s'il s'agissait d'enregistrements fabriqués, aucun répondant ne pourrait prétendre connaître les personnes en question.

Au moment de l'expérimentation du nouveau plan de sondage dans le TOR, les intervieweurs affectés à l'échantillon D devaient déterminer si des enregistrements de cet échantillon étaient fictifs et devaient indiquer les motifs de leurs décisions. À l'origine, les commis exigeaient des preuves solides avant de reconnaître une personne de l'échantillon D comme fictive. Ces données ont servi à établir les premières estimations de la population totale et du taux de sous-dénombrement selon le TOR. Nous nous sommes rendu compte plus tard que les règles de codage étaient interprétées trop rigoureusement et nous avons demandé à des spécialistes de réexaminer tous les cas de l'échantillon D qui avaient été classés parmi les cas de "non-interview" (inconnu du répondant) pour déterminer s'il ne s'agissait pas plutôt de cas "fictifs". Sur 257 cas réexaminés, 118 ont été reconnus comme fictifs par les spécialistes. Les données corrigées ont servi à établir de nouvelles estimations pour le TOR (Schenker 1988).

En règle générale, le codage géographique des questionnaires du recensement a été reconnu comme très bon dans la région d'essai de Los Angeles, qui était un ancien quartier formé de grands blocs bien circonscrits. Nous n'avons pas évalué formellement l'effet des erreurs de codage géographique sur le nombre estimé d'enregistrements erronés mais nous croyons que ces erreurs sont négligeables. Dans d'autres régions des E.-U., toutefois, ce genre d'erreur pourrait ne pas être négligeable en raison de cartes de mauvaise qualité, d'adresses incomplètes ou d'une confusion à propos de la position géographique créée par de nouvelles constructions. Par exemple, contrairement au test de Los Angeles, le test du Mississippi pour 1986 a donné lieu à des problèmes de codage géographique. Nous avons constaté dans ce test que 2,22% des enregistrements de l'échantillon D étaient répétés. De ce groupe, 35% se trouvaient à l'extérieur du bloc échantillonné.

Bien que nous ayons pu trouver de nombreux enregistrements répétés à l'extérieur de l'échantillon, nous ne sommes pas sûrs de les avoir tous trouvés car la recherche de tels enregistrements ne constituait pas une activité distincte. Nous nous sommes contentés de relever

pour l'ensemble de l'échantillon D dans le TOR est égal au taux observé dans l'EP de 1980; ceci est toutefois attribuable à une erreur opérationnelle dans le TOR et des réductions des taux de données manquantes devraient, dans l'avenir, être aussi prononcées que celles observées dans le cas de l'échantillon P.

Même si nous avons obtenu des taux de données manquantes relativement faibles dans le TOR, il importe d'analyser l'effet des données manquantes sur les taux de sous-dénombrement estimés. À cette fin, nous avons calculé plusieurs séries d'estimations du taux de sous-dénombrement pour le TOR au moyen de diverses méthodes de traitement des données manquantes, des cas d'interview par procuration dans l'échantillon P, des cas de démenagement dans l'échantillon P et de certains cas non résolus de l'échantillon D. Voir Schenker (1988) pour une description détaillée des diverses estimations, qui varient de 7,8% à 9,4%. Deux des méthodes de traitement considérées dans Schenker (1988) visent à résoudre des problèmes que nous analysons ailleurs dans cet article; il s'agit du traitement des cas résolus de l'échantillon sans quitter la région d'essai (sections 2 et 3) et du traitement des cas résolus de l'échantillon D qui peuvent constituer des enregistrements fictifs (section 6). Dans le tableau 1, les effets de ces méthodes de traitement sont attribués à d'autres sources d'erreur que les données manquantes et ils expliquent en majeure partie l'écart entre le taux de sous-dénombrement estimé du TOR (9%) et la plus faible estimation obtenue par Schenker (7,8%). Lorsque nous considérons les autres méthodes de traitement analysées dans Schenker (1988), la différence dans le taux de sous-dénombrement estimé varie de -0,3% à 0,3%. Il s'agit de différences relativement faibles pour lesquelles il est difficile de dire dans quel sens est orienté l'effet. C'est pour quoi nous avons indiqué dans le tableau 1 un effet moyen de 0,0% pour les données manquantes.

6. ERREUR DANS L'ESTIMATION DU NOMBRE D'ENREGISTREMENTS ERRONÉS

Pour estimer l'erreur de couverture nette, il faut estimer le nombre d'enregistrements erronés (EE) créés lors du recensement. Les EE comprennent les catégories suivantes:

- i) fabrication dans le recensement, par laquelle le recenseur invente des personnes au lieu de réaliser une interview en bonne et due forme ou par laquelle le répondant fournit le nom de personnes fictives;
- ii) répétition d'enregistrements;
- iii) personnes nées après le jour du recensement et personnes décédées avant ce jour;
- iv) personnes recensées avec si peu de renseignements qu'il n'est pas possible de les rattacher aux enregistrements de l'EP.

Toutes ces catégories sont estimées au moyen de l'échantillon D. En outre, certaines erreurs de géocodage dans le recensement sont considérées comme des enregistrements erronés; cette question s'inscrit dans le problème de l'équilibrage analysé dans la Section 7.

Dans l'EP de 1980, l'échantillon D était un échantillon distinct et indépendant de 110,000 ménages recensés. Les intervieweurs ont visité de nouveau les unités de logement huit mois après le jour du recensement pour vérifier si les enregistrements d'alors étaient corrects ou erronés. De plus, on a situé chaque unité de logement sur une carte pour vérifier si on lui avait attribué le bon code géographique et des commis ont examiné les dossiers du recensement pour déceler les enregistrements répétés.

Depuis 1980, l'échantillon D a fait l'objet de deux modifications majeures. Premièrement, comme nous l'avons déjà mentionné, les échantillons P et D reposent désormais sur le même échantillon de blocs. Nous avons constaté que le chevauchement des échantillons P et D a pour effet de réduire les erreurs de codage géographique. En deuxième lieu, la plupart des données de l'échantillon D seront recueillies en juillet, soit à peine trois mois suivant le jour du recensement. Désormais, on peut dire qu'une personne de l'échantillon D a été enregistrée correctement si elle est comptée dans l'échantillon P en juillet et qu'elle peut ensuite être apparée

Tableau 5

Taux de données manquantes des EP (%)^a

Source	EP de 1980		TOR de 1986
	Avril	Août	
Échantillon P	Non-interview (ménage)	4.4	0.5
	Code de dénombrement indéterminé		
	(personne)	4.0	0.8
	Total	8.4	1.3
	Interview par procuration (ménage)	a	3.2
Échantillon D	Non-interview (ménage)	1.1	SO
	Code géographique indéterminé		
	(ménage)	1.6	SO
	Code de dénombrement indéterminé		
	(personne)	2.0	4.7
Total			4.7

^a Pourcentage inconnu.

NOTA: S.O. signifie "sans objet".

Premièrement, en raison du délai serré prévu pour les interviews de la CPS, les interviews initiales pour l'échantillon P de 1980 ont été réalisées sur une période d'une semaine. Le nouveau plan de sondage pour l'EP prévoit une période d'interview de trois semaines plus une semaine additionnelle s'il survient des difficultés particulières. La prolongation de la période d'interview a pour effet de réduire le taux de non-interview des ménages. On a également réussi à réduire le taux de non-interview des ménages en utilisant un échantillon de blocs de recensement au lieu des grappes de quatre unités de logement tirées d'une liste comme dans le cas de la CPS. L'échantillon de blocs permet à l'intervieweur de visiter une unité de logement à plusieurs reprises (peut-être entre deux visites d'unités de logement situées dans le même bloc) sans que cela n'entraîne des frais de déplacement trop élevés.

Les interviews de rappel incomplètes expliquent une bonne partie des codes de dénombrement manquants pour l'échantillon P dans l'EP de 1980 (2.6% pour avril et 2.8% pour août). Nous tentons actuellement d'améliorer ce problème en recueillant durant l'interview initiale les données qui permettront de déterminer si une personne a été recensée ou manquée, éliminant ainsi, dans la plupart des cas, la nécessité d'un suivi. En outre, l'amélioration de la rapidité et de la qualité de l'appariement, grâce au nouveau système d'appariement informatisé, aura pour effet de réduire le nombre de cas exigeant un suivi.

Selon le nouveau plan de sondage de l'EP, les échantillons P et D se chevauchent, ce qui fait que la plupart des renseignements nécessaires à la détermination des codes de dénombrement pour l'échantillon D sont recueillis tôt dans le processus, soit durant l'interview initiale de l'échantillon P. L'utilisation d'un échantillon de blocs et l'application d'une géographie du recensement améliorée permettent aussi de réduire la proportion de cas de l'échantillon D pour lesquels on ne peut évaluer le degré d'exactitude du géocodage de recensement. Enfin, des améliorations ont été apportées au traitement des données manquantes (Schenker 1988). Comme l'indique le tableau 5, les taux de données manquantes pour l'échantillon P dans le TOR sont très inférieurs aux taux observés dans l'EP de 1980. Le taux de données manquantes

Le nouveau plan de sondage de l'EP a été conçu de manière à réduire au minimum le taux de fabrication. Il prévoit des contrôles qualitatifs fréquents fondés sur la réinterview. Plusieurs fois par semaine, on prélève des échantillons dans chaque bloc pour vérifier le travail de chaque intervieweur. Une vérification aussi serrée n'était pas possible dans l'EP de 1980 puisque la tâche des intervieweurs était moins concentrée géographiquement. En outre, la formation et la supervision des intervieweurs ont aussi été améliorées depuis ce temps. On a mis en place des programmes de recyclage et des mécanismes de rétroaction pour faire en sorte que les intervieweurs ne répètent pas leurs erreurs.

Deux études ont permis d'évaluer l'importance de la fabrication dans le TOR de 1986. En premier lieu, des contrôles qualitatifs rigoureux ont été effectués pendant la collecte des données pour l'échantillon P; ces contrôles portaient aussi bien sur le listage d'adresses que sur l'interview. Essentiellement, ils n'ont révélé l'existence que d'un petit nombre d'enregistrements fabriqués. Quelques jours après l'interview initiale de l'échantillon P, des commis ont soumis les résultats de 2,070 interviews à un contrôle qualitatif consistant à vérifier la composition des ménages. Sur ces 2,070 interviews, 59 n'ont pas satisfait aux critères du contrôle qualitatif. On a analysé minutieusement ces cas pour déterminer combien d'entre eux étaient le produit d'une fabrication. À cette fin, on a tenté d'apparier chaque membre du ménage à l'enregistreur du recensement en se fondant sur les données fournies par le premier intervieweur (et non sur celles fournies par le commis affecté au contrôle qualitatif); une concordance impliquerait que le premier intervieweur avait recueilli des données justes au sujet de cette personne. Pour invalider cette hypothèse, il faudrait qu'il y ait eu une fabrication identique au moment du recensement. Seulement 13 des 59 cas ont été jugés comme des cas possibles de fabrication; par exemple, aucun membre du ménage figurant sur la liste originale de l'EP n'avait pu être apparié à un enregistrement du recensement. Par conséquent, le taux de fabrication estimé pour le contrôle qualitatif était de 0.6%.

La seconde source de données nous permettant d'évaluer l'importance de la fabrication est le suivi effectué après production des estimations initiales, qui a été décrit dans la section 3. D'après les données du tableau 4, nous estimons qu'environ 1.2% des personnes de l'échantillon P peuvent avoir été le fruit d'interviews fictives. Ce taux de fabrication est environ le double de celui estimé par suite du contrôle qualitatif. Nous croyons qu'une bonne partie de la différence est imputable à un mauvais intervieweur dont les actions ont été dévoilées par l'interview de rappel, après avoir manifestement échappé au contrôle qualitatif. La différence de taux de fabrication peut-être aussi par le fait que le suivi exagère le degré de fabrication; en effet, des propriétaires et d'autres répondants peuvent nier l'existence de personnes qui occupent des unités de logement transformées sans autorisation ou qui sont au pays sans statut légal. Afin de déterminer une limite supérieure pour l'effet de la fabrication dans le TOR, nous utilisons le taux de fabrication le plus élevé (.012) et nous supposons que le taux d'appariement pour les personnes de l'échantillon P qui auraient été identifiées à l'aide d'interviews convenables est le même que celui observé pour les cas non fabriqués, soit environ .88. Ces hypothèses permettent de ramener le taux de sous-dénombrement à environ 7.9%, soit à environ 1.1% de moins que le taux initial de 9%. Si on utilisait le taux de fabrication le moins élevé (.006), on obtiendrait par les mêmes calculs un taux de sous-dénombrement de 8.4%, soit environ .6% de moins que le taux enregistré dans le TOR. Le tableau 1 indiquait un effet de 1% pour la fabrication, ce qui correspond à peu près à la limite supérieure calculée ci-dessus.

5. DONNÉES MANQUANTES

Si nous voulons mesurer correctement les faibles erreurs de couverture, l'EP doit produire un ensemble de données aussi complet que possible, qui ne présente pas une forte proportion de données manquantes. Malheureusement, l'enquête de 1980 était caractérisée par une très forte proportion de données manquantes (U.S. Bureau of the Census 1988). Les modifications apportées au plan de sondage pour l'EP devraient désormais permettre de réduire cette proportion.

Tableau 4

Résultats du suivi effectué après production des estimations initiales, personnes (non-pondérées)

Résultat	Non-concordance totale		Concordance partielle		Concordance totale	
	avec élément de contradiction	sans élément de contradiction	Personnes non apparées	Personnes apparées	du ménage	du ménage
#		#	#	#	#	#
%		%	%	%	%	%
Adresse confirmée	64	33	252	73	61	75
Nouvelle adresse donnée	32	17	46	13	13	16
Fabrication possible	70	36	23	7	2	2
Non-interview	27	14	24	7	5	6
Total	193	100	345	100	81	100
					153	100
					165	100

Nota: Le symbole # signifie le nombre de personnes faisant partie du sous-échantillon utilisé pour le suivi. Le symbole % signifie le pourcentage du total de colonne.

Certains renseignements nous permettent de croire que 95 cas auraient possiblement inventés lors de l'interview initiale de l'échantillon P. La plupart de ces cas (70) se trouvent dans la catégorie "non-concordance totale du ménage avec élément de contradiction". Ce problème est traité en détail dans la section 4. Par ailleurs, il y a des cas où la réinterview n'a pas été réalisée en entier ou n'a pas produit suffisamment de données pour que l'on puisse classer les personnes dans l'une ou l'autre catégorie. Si l'interview avait été menée correctement, on aurait peut-être relevé une nouvelle adresse pour certains de ces cas.

Si nous pondérons les données du tableau 4 aux totaux de l'échantillon P, nous estimons qu'environ 3,1% des personnes de cet échantillon ont été considérées par erreur comme des personnes n'ayant pas déménagé selon l'interview initiale. En ce qui concerne les personnes qui ont déménagé sans quitter la région d'essai, nous avons pu tenter un appariement à la nouvelle adresse et constaté qu'un tiers des cas avaient été dénombrés lors du recensement d'essai de Los Angeles. Pour évaluer l'effet probable des erreurs de déclaration, surtout si nous considérons le TOR comme le test d'une EP menée à l'échelle nationale, nous supposons que les personnes qui ont indiqué des adresses à l'extérieur de la région d'essai auraient été recensées dans la même proportion que les personnes qui ont indiqué des adresses dans la région d'essai. Ainsi, le tiers des personnes qui ont été considérées par erreur comme des personnes n'ayant pas déménagé auraient été apparées et classées parmi les personnes recensées. Le redressement des estimations pour l'erreur de déclaration a pour effet de réduire de un pourcent l'estimation du taux de sous-dénombrement.

4. FABRICATION DANS L'INTERVIEW DE L'EP

Malgré tout l'effort que l'on met à former et à suivre les intervieweurs, un intervieweur pour l'EP peut parfois inventer un ménage au lieu de réaliser une interview en bonne et due forme. Les enregistrements fabriqués ne peuvent être apparés aux enregistrements du recensement. Ce phénomène contribuera à gonfler l'estimation du taux de sous-dénombrement dans la mesure où les enregistrements fabriqués pour une adresse donnée tiendront la place de personnes qui ont effectivement été recensées.

Tableau 3
Tailles d'échantillons pour le suivi effectué après la
production des estimations initiales

Résultat de l'appariement initial		Concordance totale du ménage	
		dans l'échan- tillon P	réin- tervues
Concordance totale du ménage		4,662	50
Concordance partielle du ménage		609	50
Non-concordance totale avec élément		160	64
Non-concordance totale sans élément		357	109

suivait de six mois l'interview initiale de l'EP et de dix mois le recensement. Avant d'exposer les résultats de cette étude, nous signalons ici deux contraintes qu'il faut considérer. La première est le risque d'une plus grande erreur de rappel que dans l'interview originale. La seconde est la possibilité d'effritement de la confiance qu'avait inspirée la campagne de publicité pour le recensement, cette contrainte pouvant constituer un problème sérieux dans les régions qui comptent un grand nombre d'immigrants sans statut légal, qui craignent tout contact avec des représentants de l'Etat.

Le tableau 3 décrit la composition du sous-échantillon. Dans la plupart des cas, le ménage de l'EP concorde parfaitement avec celui du recensement (concordance totale). Pour les cas de concordance partielle, quelques-uns des membres du ménage de l'EP concordent avec l'entre-gistement du recensement, d'autres non.

La catégorie "non-concordance totale du ménage avec élément de contradiction" constitue ce que nous appelons le problème "Dufour-Paradis". La famille Dufour a été recensée à une certaine adresse et le suivi exécuté pour l'échantillon D a confirmé cette adresse. Or, l'interview réalisée pour l'échantillon P a révélé qu'une famille Paradis demeurait à cette adresse le jour du recensement. Il y a là une contradiction et elle s'explique peut-être par le fait que les Paradis ont mal indiqué l'adresse où ils demeureraient le jour du recensement. La catégorie "non-concordance totale du ménage sans élément de contradiction" ne renferme pas de contradictions apparentes; par exemple, l'unité de logement peut avoir été oubliée lors du recensement ou classée dans les logements inoccupés.

Le tableau 4 donne les résultats pour les personnes de l'échantillon en distinguant les personnes apparues initialement des personnes non apparues initialement dans les cas de concordance partielle. Comme prévu, le pourcentage de personnes pour lesquelles l'adresse a été confirmée varie considérablement d'une strate à l'autre. Les adresses ont pratiquement toutes été confirmées en ce qui a trait à la catégorie "concordance totale" tandis que le taux de confirmation le plus faible a été enregistré pour la catégorie "non-concordance totale du ménage avec élément de contradiction". De 13 à 17% des personnes non apparues, selon l'une ou l'autre des trois catégories concernées, ont donné de nouvelles adresses. Chose intéressante, on a indiqué de nouvelles adresses pour 10% des personnes apparues qui sont membres de ménages partiellement apparus, ce qui n'est pas beaucoup moins que le pourcentage correspondant pour les personnes non apparues des mêmes ménages. Il y a peu de chances que la nouvelle adresse soit exacte à moins que l'on ait commis la même erreur lors du recensement et dans l'interview initiale de l'échantillon P. La variabilité des résultats observés nous confirme dans notre opinion qu'un suivi réalisée plusieurs mois après l'interview initiale de l'échantillon P se solde par-fois par des réponses différentes (à cause de la crainte et des erreurs de rappel) mais ne permet pas nécessairement d'obtenir une adresse plus juste.

Tableau 2

Résultats de l'étude sur le nouvel appartement: échantillon (pondéré)^a

Résultats de l'appartement initial	Résultats du nouvel appartement			
	Recensés	Non recensés	Non résolus	Total
Recensés	16,623	18	55	16,696
Non recensés	88	2,164	56	2,308
Non résolus	17	0	132	149
Total	16,728	2,182	243	19,153

^a Les données sont pondérées aux totaux de l'échantillon P.

La seconde étude visait à évaluer l'importance de l'erreur d'appariement pour les personnes ayant déménagé. Parmi les personnes qui n'ont pu être appariées dans le TOR, 90 ont déclaré qu'elles avaient déménagé entre le jour du recensement et le moment où s'est déroulée l'EP. Pour les personnes ayant déménagé, la recherche se fait en fonction de l'adresse déclarée le jour du recensement. Pour évaluer la qualité du processus d'appariement, nous avons réexaminé les 90 cas de déménagement à l'aide de méthodes plus rigoureuses. Cet exercice nous a permis de découvrir 11 nouveaux cas de concordance, ce qui a fait passer le taux d'appariement observé pour les personnes ayant déménagé et faisant partie du champ de l'enquête de .661 à .719, pour une augmentation de .058. Bien que le taux de "fausses non-concordances" ($(11/90) = .122$) pour les personnes ayant déménagé soit supérieur à celui observé pour les personnes n'ayant pas déménagé, le premier groupe de personnes représente une proportion relativement faible de l'échantillon P. Si nous corrigeons le biais par défaut dont est entaché le taux d'appariement pour les personnes ayant déménagé et celles n'ayant pas déménagé (5.8 et 0.6% respectivement), nous nous trouvons à réduire de 0.7% le taux de sous-dénombrement du TOR.

Ces calculs ne tiennent pas compte de la possibilité de nouveaux cas de concordance dans l'hypothèse où l'objet de la première étude irait au-delà des limites des blocs visés par l'EP (Thompson, Whitford et Stoudt 1987). Toutefois, si nous nous fondons sur les résultats de l'appariement informatisé pour la région d'essai de Los Angeles, nous en concluons que le codage géographique a été fait correctement et que l'observation de nouveaux cas de concordance pourrait entraîner une baisse additionnelle du taux de sous-dénombrement estimé du TOR d'au plus 0.3%.

3. DÉCLARATION DE L'ADRESSE LE JOUR DU RECENSEMENT

Dans le nouveau plan de sondage de l'EP, nous tentons, comme dans le plan de 1980, d'appari-er les personnes de l'échantillon P à l'enregistrement qui correspond à l'adresse le jour du recensement. Pour faciliter l'appariement, l'intervieweur doit demander à quel endroit vivait chaque membre du ménage le jour du recensement. Il cherche ensuite à savoir si ces personnes auraient pu demeurer à une autre adresse ce même jour, par exemple sur un campus de collège ou d'université, sur une base militaire ou à bord d'un bâtiment de la marine ou encore dans une résidence secondaire. Si l'adresse le jour du recensement déclarée incorrectement lors de l'interview de l'échantillon P, il y a alors risque de classer par erreur les membres de ce ménage parmi les personnes non recensées, ce qui aurait pour effet d'introduire un biais par excès dans l'estimation du taux de sous-dénombrement.

Pour évaluer l'importance de l'erreur de déclaration d'adresses, nous avons réinterviewé un sous-échantillon de personnes appariées et des personnes non appariées après que les esti-mations initiales du taux de sous-dénombrement aient été produites. Cette interview de rappel

Cette classification, qui n'était pas possible selon le plan de 1980, rend le processus d'appariement plus fiable. Par exemple, on peut désormais distinguer les personnes qui ont le même nom dans un quartier ethnique en utilisant tous les renseignements tirés d'un échantillon de blocs. De même, l'échantillon de blocs facilite le débrouillement des erreurs d'adresses. Le choix du bloc de recensement comme unité d'échantillonnage a aussi pour effet de réduire les erreurs de géocodage par rapport à l'EP de 1980, où l'échantillon P était fondé sur des grappes de quatre unités de logement de la CPS et sur la géographie du recensement de 1970.

L'appariement est particulièrement difficile pour les personnes de l'échantillon P qui demeuraient à une autre adresse le jour du recensement (personnes ayant déménagé). Pour cette catégorie de personnes, il est indispensable d'associer l'adresse le jour du recensement déclarée lors de l'EP à la région géographique appropriée avant de procéder à l'appariement. Cette opération posait des problèmes dans l'EP de 1980 et ces problèmes ne sont pas nécessairement résolus avec le nouveau plan de sondage. Toutefois, le Bureau of the Census mettra en application un nouveau système géographique automatisé pour le recensement de 1990 (voir Marx et Saalfeld 1988) et nous espérons que cette innovation permettra d'associer avec célérité et justesse l'adresse des personnes ayant déménagé à la région géographique appropriée.

Lors du TOR de 1986, environ 74% des personnes de l'échantillon P ont été appariées par ordinateur. En outre, 12% des membres de ce même échantillon ont été reconnus comme des cas de "concordance probable" par l'ordinateur. Des commis formés spécialement à cette fin ont passé en revue tous les cas que l'ordinateur n'a pas reconnus comme des cas de "concordance", y compris tous les cas de "concordance probable".

Les résultats de l'EP de 1986 au Mississippi montrent que l'efficacité du système d'appariement informatisé ne se limite pas aux régions urbaines, caractérisées par des numéros civiques, des noms de rue et une géographie bien définie. Dans le test effectué au Mississippi, les adresses étaient composées le plus souvent d'une route rurale et d'un numéro de case postale. Les blocs, de forme irrégulière, étaient définis par des limites invisibles comme un cours d'eau intermittent ou la frontière d'un comté. Néanmoins, l'ordinateur a pu faire l'appariement pour 68% des personnes.

Nous avons réalisé deux études dans le but d'évaluer l'importance de l'erreur d'appariement dans le TOR. Dans la première étude, des spécialistes du bureau chef ont prélevé un sous-échantillon de 35 blocs et procédé à un nouvel appariement. On a effectué cette opération indépendamment de l'appariement initial, puis on a déterminé quels étaient les écarts entre les résultats des deux opérations. À cause de la rigueur avec laquelle s'est fait le nouvel appariement, nous croyons que les résultats de cet appariement reflètent la réalité tandis que les écarts entre les résultats du premier et du second appariement reflètent le biais dont sont entachés les résultats du premier appariement. Seules les personnes n'ayant pas déménagé ont été considérées. En outre, l'étude s'est limitée à un nouvel appariement à l'intérieur des blocs, de sorte qu'elle n'a pas permis d'évaluer avec exactitude le nombre de "fausses non-concordances" qui pouvaient être attribuables au fait que l'enregistrement du recensement se trouvait à l'extérieur du bloc visé par l'EP.

Le tableau 2 expose les résultats de l'étude pour l'échantillon P sous la forme d'une table où se retrouvent les codes d'appariement obtenus à la suite de l'appariement initial du TOR et ceux obtenus à la suite du nouvel appariement.

On estime que les résultats du TOR renferment environ 88 fausses non-concordances et 18 fausses concordances et que 111 (55 + 56) cas qui ont été classés initialement comme appariés ou non appariés auraient dû être reconnus comme non résolus. Dans le cours normal de l'estimation, les cas non résolus devraient être traités au moyen des méthodes appliquées aux données manquantes (voir Schenker 1988). Il ressort du tableau 2 que le taux d'appariement observé (c'est-à-dire le nombre de cas appariés divisé par le total des cas appariés et des cas non appariés) est de .879 pour l'appariement initial et de .885 pour le nouvel appariement et que, par conséquent, le taux d'appariement initial est entaché d'un biais par défaut d'environ 0.6%.

Par conséquent, si nous devons tenir compte de l'effet combiné des erreurs dans l'estimation du taux de sous-dénombrement, celui-ci serait ramené de 9,0% à environ 7,8%. Le chiffre redressé (7,8%) peut être considéré à peu près comme la moyenne d'une distribution a posteriori de l'erreur pour le taux de sous-dénombrement déterminé à l'aide du TOR. Le Bureau of the Census s'emploie actuellement à élaborer une distribution a posteriori de l'erreur complète (voir Mulry et Spencer 1988). Comme l'estimation initiale du TOR (9%) est beaucoup plus près du pourcentage redressé (7,8%) que celui-ci ne l'est de zéro, nous en concluons que les données originales du TOR reflètent mieux la réalité que les données du recensement.

Dans les huit sections qui vont suivre, nous allons examiner les catégories d'erreur une par une. Chaque section expose les méthodes utilisées dans l'EP de 1980 et les problèmes qui y étaient rattachés, de même que les perfectionnements permettant de réduire la marge d'erreur qui ont été éprouvés dans le TOR. En outre, nous décrivons le mode d'évaluation de chaque source d'erreur et présentons les résultats qui ont été à l'origine de nos conclusions. Finalement, la section 10 présente un résumé des résultats de notre analyse et propose certaines orientations pour la recherche future.

2. ERREUR D'APPARIEMENT

Deux raisons générales peuvent expliquer une mauvaise classification des membres de l'échantillon P :

a) soit que l'information fournie par le répondant ou l'intervieweur est inexacte;

b) soit que l'information fournie est exacte mais utilisée incorrectement.

La première catégorie concerne l'enregistrement de l'adresse déclarée le jour du recensement et la fabrication d'enregistrements dans l'EP, deux sujets traités dans les sections 3 et 4 respectivement. La présente section porte sur les erreurs d'appariement (catégorie b), qui surviennent même lorsqu'il s'agit de personnes véritables et que l'adresse le jour du recensement a été déclarée correctement. Autrement dit, nous parlons ici des erreurs d'appariement attribuables à des erreurs de traitement.

Dans le nouveau plan de sondage de l'EP, l'appariement revêt deux formes : appariement automatisé par lots et appariement manuel assisté par ordinateur. Une personne de l'échantillon P est reconnue comme "non recensée" lorsqu'on a recueilli une quantité suffisante de données pour l'appariement mais qu'il n'est pas possible d'apparier la personne en question à un enregistrement du recensement. Des erreurs surviennent lorsqu'on ne dispose pas de données suffisantes pour l'appariement et que l'on tente néanmoins d'apparier la personne en question ou lorsqu'on recherche parmi les bons questionnaires de recensement sans pouvoir établir une concordance même si la personne a effectivement été recensée.

Il arrive qu'une personne de l'échantillon P soit apparée au mauvais enregistrement. Cela se produit le plus souvent à l'intérieur d'une même famille, où des enfants peuvent avoir des noms et des âges semblables, et dans les quartiers "ethniques", où certains noms sont anormalement répandus. En règle générale, les fausses concordances sont plus rares que les fausses non-concordances car les cas de concordance peuvent être vérifiés facilement par des commis posés à l'appariement manuel. Les fausses concordances introduisent un biais dans l'estimateur bivalent seulement si le membre de l'échantillon P n'a effectivement pas été recensé. Un des principaux changements au plan de sondage de l'EP depuis 1980 est l'utilisation d'un même échantillon de blocs pour les échantillons P et D ; cela permet un meilleur contrôle de l'erreur d'appariement. L'échantillonnage de blocs permet de classer toutes les personnes enquêtées (échantillons P et D) en trois catégories :

- incluses dans les échantillons P et D
- incluses dans l'échantillon P mais non dans l'échantillon D
- incluses dans l'échantillon D mais non dans l'échantillon P.

Erreurs du TOR et estimations de l'effet moyen de la correction
d'erreur sur le taux de sous-dénombrement estimé

Sources d'erreur		Effet moyen sur le taux de sous-dénombrement estimé
<hr/>		
Erreur d'appariement	- 1.0%	
Déclaration de l'adresse le jour du recensement	- 1.0%	
Fabrication dans l'interview de l'EP	- 1.0%	
Données manquantes	0.0%	
Erreur dans l'estimation du nombre d'enregistrements erronés	- 0.5%	
Conciliation des taux de sur-dénombrement et de sous-dénombrement bruts	0.0%	
Biais de corrélation	+ 2.3%	
Erreur aléatoire	0.0%	

logement et 20,000 personnes. Ce test a permis d'établir un taux de sous-dénombrement net estimé d'environ 9%. Pour plus de détails sur les méthodes et les résultats du TOR, voir Dif-fendal (1988) et Schenker (1988).

Nous avons aussi expérimenté le nouveau plan de sondage dans une région rurale du Mis-sissippi la même année. Nous avons utilisé à cette fin un échantillon de 271 blocs compren-nant environ 3,250 unités de logement et 8,000 personnes. Ce test a permis d'établir un taux de sous-dénombrement estimé de 5.5%. Pour plus de détails sur les résultats et les méthodes utilisées, voir Anolik (1988). Bien que les données de ce dernier test n'aient pas été analysées aussi minutieusement que les données du TOR, nous nous reporterons parfois aux résultats de l'analyse au cours de cet article.

Une question importante est de savoir si la nouvelle EP peut produire des estimations démogra-phiques plus justes que ne peut le faire le recensement. En théorie, les estimations de l'EP devraient être considérées comme les plus justes mais, en pratique, la réalisation de l'EP et du recensement, de même que l'analyse des résultats, ne sont pas à l'abri des erreurs non dues à l'échantillonnage. Une analyse sérieuse est nécessaire pour évaluer la justesse relative de ces deux types d'enquête. Dans cet article, nous nous attachons à évaluer la structure d'erreur du TOR de 1986.

Huit sources d'erreur peuvent influencer sur les estimations du taux de sous-dénombrement produites par l'EP: l'erreur d'échantillonnage plus sept sources d'erreur non due à l'échantil-lonnage. Le tableau 1 présente ces sources d'erreur ainsi qu'une évaluation sommaire de leur incidence sur les données du TOR. La seconde colonne du tableau indique l'effet des erreurs sur le taux de sous-dénombrement estimé. Par exemple, si nous corrigeons toutes les "erreurs d'appariement", le taux de sous-dénombrement estimé baissera d'environ un point de pour-centage, soit de 9 à 8%. Certaines catégories d'erreur, comme les "données manquantes" et l'"erreur aléatoire", peuvent contribuer à accroître ou à diminuer le taux de sous-dénombrement mais nous croyons qu'elles n'introduisent pas de biais majeur dans les don-nées du TOR. Les chiffres de la seconde colonne ont été déterminés uniquement en fonction de la source d'erreur correspondante (c'est-à-dire que l'on n'a pas tenu compte de l'effet que pouvaient avoir concurremment les autres sources d'erreur).

Par construction, les huit catégories d'erreur tendent à s'exclure mutuellement et à s'addi-tionner. Il peut y avoir des chevauchements ou des interactions entre les diverses catégories, mais nous croyons qu'ils sont négligeables et nous n'en tenons pas compte ici. Globalement, nous calculons l'effet combiné des erreurs comme suit:

$$(-1.0 - 1.0 - 1.0 - 1.0 + 0.0 - 0.5 + 0.0 + 2.3 + 0.0) \text{ pour cent} = -1.2 \text{ pour cent.}$$

indépendant constitué de personnes qui vivaient dans des unités de logement recensées. En outre, le Bureau of the Census avait produit une autre série d'estimations du taux de sous-dénombrement fondées sur une analyse globale des registres des naissances et des décès, de fichiers administratifs et de recensements antérieurs. Ce programme, appelé analyse démographique, sera évoqué à quelques reprises dans notre étude. Le Bureau of the Census n'a pas redressé les chiffres du recensement de 1980 en fonction de l'erreur de couverture car il considérait que les estimations de l'EP étaient faussées par des données manquantes ou inexacts. En outre, les résultats de l'analyse démographique laissaient à désirer du fait, notamment, que l'on ne disposait pas de données suffisantes sur le nombre d'immigrants sans statut légal et qu'il n'existait pas de méthode satisfaisante pour établir des estimations au niveau de l'État ou de la municipalité. Voir U.S. Bureau of the Census (1988).

Ces dernières années, nous avons élaboré pour l'EP un nouveau plan de sondage et de nouvelles méthodes qui atténuent les problèmes que nous avons connus en 1980 sans en créer d'autres d'importance majeure. Le nouveau plan de sondage de l'EP est fondé sur un échantillon aréolaire de blocs de recensement à partir duquel est constitué l'échantillon P. L'échantillon P comprend toutes les personnes qui demeurent dans les blocs échantillonnés au moment de l'interview de l'EP. Les intervieweurs vont dans chaque unité de logement et déterminent à quel endroit vivaient les occupants au moment du recensement.

À l'aide d'un logiciel et de méthodes d'appariement élaborées récemment (Jaro 1988), nous tentons d'apparier toutes les personnes de l'échantillon P aux enregistrements du recensement correspondants. Des commis vérifient les résultats de l'appariement informatisé et attribuent à chaque personne de l'échantillon P un code de dénombrement qui indique si la personne a été enregistrée ou manquée lors du dénombrement initial. Pour ce qui a trait aux personnes qui ont déménagé entre le jour du recensement et le jour de l'EP, nous associons l'adresse déclarée le jour du recensement au bloc approprié et tentons d'établir l'appariement à cet endroit. Il arrive que l'appariement ne soit pas décisif pour tous; pour les quelques cas en suspens, une interview de rappel est alors nécessaire pour recueillir des renseignements additionnels ou résoudre les incohérences que renferment les données déjà recueillies. Après le suivi, les commis attribuent un code de dénombrement aux personnes de l'échantillon P pour lesquelles on a complété une interview de rappel. Malgré toutes ces étapes, il est possible qu'un très petit nombre de cas demeurent en suspens; nous imputons alors à chacun d'eux un code de dénombrement en appliquant des méthodes statistiques appropriées pour les données manquantes (Schenker 1988).

En ce qui a trait aux personnes de l'échantillon D, on attribue également à chacune d'elles un code de dénombrement indiquant si la personne a été enregistrée correctement ou incorrectement lors du recensement. Dans la section 6, nous décrivons ce qui constitue un enregistrement incorrect ou erroné (BE); par ailleurs, tous les enregistrements non erronés sont considérés comme des enregistrements corrects (EC). Un grand nombre des personnes interviewées comme membres de l'échantillon P ont aussi été enregistrées lors du recensement. Les deux échantillons se chevauchent donc dans une forte proportion. Tous les membres de l'échantillon D qui font aussi partie de l'échantillon P (conformément aux résultats de l'appariement informatisé et manuel) sont reconnus automatiquement comme des enregistrements corrects (EC). Toutefois, le chevauchement n'est pas parfait. L'échantillon D pourra compter des personnes qui ne figurent pas dans l'échantillon P et vice versa. D'autres personnes dans le bloc pourront être recensées par erreur. Des intervieweurs pourront fabriquer des enregistrements. Chaque personne de l'échantillon D qui n'aura pu être appariée à une personne de l'échantillon P devra faire l'objet d'une interview de rappel. Cette interview permet de recueillir suffisamment de données pour déterminer si une personne de l'échantillon D a été enregistrée correctement lors du recensement.

Le nouveau plan de sondage de l'EP a été expérimenté en 1986 à l'occasion d'un recensement d'essai réalisé à Los Angeles. Ce recensement, qui a été désigné comme le Test des opérations de redressement (TOR), portait sur 190 blocs renfermant près de 6,000 unités de

Mesure de l'erreur dans une enquête post-censitaire

HOWARD HOGAN et KIRK WOLTER¹

RÉSUMÉ

Le U.S. Bureau of the Census aura recours à une enquête post-censitaire pour évaluer le taux de couverture du recensement décennal de 1990. Cet article décrit les nouvelles méthodes que le Bureau a mises au point et expérimentées pour accroître le degré de précision des estimations de cette enquête. Il étudie aussi les catégories d'erreur qui surviennent dans une enquête post-censitaire de même que les moyens qui permettent de juger de l'exactitude des résultats. À cette fin, les auteurs se servent d'une enquête post-censitaire qui a été réalisée récemment à titre expérimental.

MOTS CLÉS: Recensement; sous-dénombrement; sur-dénombrement; évaluation du taux de couverture.

1. INTRODUCTION

Cet article porte sur les recherches effectuées récemment au U.S. Bureau of the Census pour accroître et mesurer la précision des estimations d'une enquête post-censitaire. À l'origine, ces recherches visaient principalement à élaborer un ensemble cohérent de principes, de méthodes et d'opérations statistiques devant permettre de redresser les chiffres du recensement des États-Unis en fonction de l'erreur de couverture. Les résultats exposés dans cet article montrent qu'il est désormais possible de produire à l'aide d'une enquête post-censitaire (EP) des estimations de la population totale qui sont plus proches de la réalité que les estimations du recensement.

Compte tenu de la décision du Département du commerce des E.-U. de ne pas redresser les chiffres du recensement de 1990 eu égard à l'erreur de couverture, les méthodes de l'EP analysées ici serviront à évaluer soigneusement le taux de couverture de ce recensement. Voir Département du commerce des E.-U. (1987). Les résultats de cette évaluation serviront ensuite à informer les utilisateurs des limites du recensement, à orienter la planification des recensements futurs et à améliorer les estimations démographiques du Bureau of the Census pour les années qui suivent l'année du recensement.

L'EP utilise deux échantillons pour mesurer l'erreur de couverture nette. D'abord, un échantillon de personnes qui devraient normalement avoir été recensées initialement est interviewé après le recensement et sert à évaluer le nombre des omissions. Nous l'appelons l'échantillon P (pour population). Il nous faut aussi un échantillon d'enregistrements du recensement pour évaluer le nombre de répétitions et les autres erreurs contenues dans les comptes du recensement. Nous l'appelons l'échantillon D (pour dénombrement). L'estimateur dual (ED) nous permet de déduire de ces échantillons une estimation de la population totale. Voir Diffendal (1988) pour une analyse détaillée des échantillons et du modèle dual. Sauf indication contraire, nous utiliserons dans cet article la notation adoptée par Diffendal.

Le Bureau of the Census a réalisé une EP en rapport avec le recensement de 1980. L'échantillon P était constitué des membres des ménages qui avaient été enquêtés lors des sondages d'avril et d'août de la Current Population Survey (CPS). Pour une description de la CPS, voir U.S. Bureau of the Census (1978). L'échantillon D était un échantillon distinct et

¹ U.S. Bureau of the Census, Washington (D.C.) 20233. Howard Hogan est chef du Undercount Research Staff. Kirk Wolter est chef de la Statistical Research Division. Cet article est un compte rendu des recherches qui ont été faites par le personnel du U.S. Bureau of the Census. Les opinions qui y sont exprimées sont celles des auteurs et ne reflètent pas nécessairement la position de l'organisme.

Tableau A2 – Suite
Résultats de la régression logistique pour l'échantillon D

Variable explicative	Codes	Coefficient estimé
Mode d'occupation	1 si logement occupé par le propriétaire, – 1 dans les autres cas	.36
Type de construction	1 si maison unifamiliale, – 1 si immeuble à logements multiples	.17
Sexe	1 pour homme, – 1 pour femme	.08
Âge 1	1 si de 0 à 14 ans, – 1 si 65 et plus, 0 dans les autres cas	– .30
Âge 2	1 si de 15 à 29 ans, – 1 si 65 ans et plus, 0 dans les autres cas	– .04
Âge 3	1 si de 30 à 44 ans, – 1 si 65 ans et plus, 0 dans les autres cas	– .34
Âge 4	1 si de 45 à 59 ans, – 1 si 65 ans et plus, 0 dans les autres cas	.10
Origine ethnique 1	1 si hispanique, – 1 si autre, 0 si asiatique	– .02
Origine ethnique 2	1 si asiatique, – 1 si autre, 0 si hispanique	– .38

BIBLIOGRAPHIE

DIFFENDAL, G. (1988). Test des opérations de redressement de 1986 dans le Central Los Angeles County, *Techniques d'enquête*, 14.

FAY, R.E., PASSEL, J.S., et ROBINSON, J.G. (1988). *The Coverage of Population and Housing Evaluation and Research Report PHC80-E4*, Washington: U.S. Census of Population and Housing.

FREEDMAN, D.A., et NAVIDI, W.C. (1986). Regression Models for Adjusting the 1980 Census, *Statistical Science*, 1, 3-39.

HOGAN, H.R., et WOLTER, K.M. (1988). Mesure de l'erreur dans une enquête post-censitaire, *Techniques d'enquête*, 14.

KROTKI, K. (1978). *Developments in Dual System Estimation of Population Size and Growth*, Edmonton, The University of Alberta Press.

MARKS, E.S., SELTZER, W., et KROTKI, K.J. (1974). *Population Growth Estimation*, New York: The Population Council.

PALMER, S. (1967). On the Character and Influence of Nonresponse in the Current Population Survey, *Proceedings of the Social Statistics Section*, American Statistical Association, 73-80.

RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

RUBIN, D.B., SCHAFER, J.L., et SCHENKER, N. (1988). Imputation Strategies for Estimating the Undercount, *Proceedings of the Fourth Annual Research Conference*, U.S. Bureau of the Census.

RUBIN, D.B., et SCHENKER, N. (1986). Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse, *Journal of the American Statistical Association*, 81, 366-374.

RUBIN, D.B., et SCHENKER, N. (1987). Logit-Based Interval Estimation for Binomial Data Using the Jeffreys Prior, *Sociological Methodology*, 17, 131-144.

SCHENKER, N. (1987). Handling Missing Data in the 1986 Test of Adjustment Related Operations, *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

WOLTER, K.M. (1986). Some Coverage Error Models for Census Data, *Journal of the American Statistical Association*, 81, 338-346.

REMERCIEMENTS

L'auteur tient à remercier Robert O'Brien, qui a calculé les estimations bivalentes utilisées dans la section 7. Il veut aussi exprimer sa reconnaissance aux arbitres pour leurs commentaires utiles et aux membres du comité d'étude sur le sous-dénombrement du Département de statistiques de l'Université Harvard pour leur participation enrichissante aux discussions sur le traitement des données manquantes dans l'estimation du sous-dénombrement.

ANNEXE

RÉSULTATS DE LA RÉGRESSION LOGISTIQUE

Tableau A1

Variable explicative	Codes	Coefficient estimé
----------------------	-------	--------------------

Ordonnée à l'origine
Code d'interview
1 si interview ordinaire, - 1 si interview par
personne interposée

Code "déménagement"
1 si la personne n'a pas déménagé, - 1 si la
personne a déménagé

Mode d'occupation
1 si logement occupé par le propriétaire, - 1
dans les autres cas

Type de construction
1 si maison unifamiliale, - 1 si immeuble à
logements multiples

Sexe
1 pour homme, - 1 pour femme

Age 1
1 si de 0 à 14 ans, - 1 si 15 et plus, 0 dans les
autres cas

Age 2
1 si de 15 à 29 ans, - 1 si 30 à 64 ans, 0 dans
les autres cas

Age 3
1 si de 30 à 44 ans, - 1 si 45 à 64 ans, 0 dans
les autres cas

Age 4
1 si de 45 à 59 ans, - 1 si 60 à 64 ans, 0 dans
les autres cas

Origine ethnique 1
1 si hispanique, - 1 si autre, 0 si asiatique
Origine ethnique 2
1 si asiatique, - 1 si autre, 0 si hispanique

Tableau A2

Résultats de la régression logistique pour l'échantillon D

Variable explicative	Codes	Coefficient estimé
----------------------	-------	--------------------

Ordonnée à l'origine
Code de questionnaire
1 si retourné par la poste, - 1 dans les autres cas

Code de pré-suivi
1 si non-concordance partielle du ménage, - 1 si
non-concordance totale

8. RÉSUMÉ ET ANALYSE

que celle utilisée dans le TOR pour les cas d'interview par personne interposée, les cas de déménagement et les cas portant le code $W1$ sont toutes inférieures aux estimations correspondantes du tableau 2 et l'écart maximum est de 0.04 point de pourcentage. Ces résultats étaient prévisibles puisque l'addition de cas ayant un taux d'appariement imputé comparable au taux d'appariement global n'est pas censée avoir beaucoup d'effet sur les estimations. Les estimations obtenues par l'application de la méthode utilisée dans le TOR sont toutes supérieures aux estimations correspondantes du tableau 2, les écarts étant de 0.34 point de pourcentage pour les personnes d'origine hispanique, de 0.50 point de pourcentage pour les personnes d'origine asiatique et de 0.38 point de pourcentage pour les autres.

Une combinaison de méthodes de pondération et d'imputation (aléatoire et non aléatoire) a servi à traiter les données manquantes dans le TOR. Les cas de non-interview dans l'échantillon P ont été résolus grâce à une méthode de redressement par pondération au niveau de l'ilot. Une méthode hot-deck a servi à imputer les caractéristiques manquantes dans les deux échantillons. Pour pallier l'absence de codes d'appariement dans l'échantillon P et de codes de dénombrement dans l'échantillon D , on a imputé des probabilités estimées par des méthodes de régression logistique.

Comme nous l'avons indiqué dans les sections 5 et 6, l'utilisation de probabilités imputées pour les codes d'appariement et de dénombrement manquants devrait faciliter le calcul de la variance due à l'imputation de ces codes. Pour évaluer entièrement cette variance, il est nécessaire de mesurer la variance due à l'estimation des paramètres de la régression logistique de même que la variance due à l'imputation étant donné β (Rubin et Schenker 1986). Il nous faut donc une matrice estimée des variances-covariances pour β . Comme on a procédé à un échantillonnage par grappes dans le TOR, l'estimation par régression logistique (section 5), qui suppose un échantillonnage aléatoire simple, ne produit pas une estimation juste de la matrice des variances-covariances. Cela ne posait pas vraiment de problèmes dans le TOR puisque celui-ci ne visait pas principalement à mesurer la variance due à l'imputation. En outre, pour ce qui a trait aux taux de non-réponse enregistrés dans le test, la variance due à l'incertitude entourant l'estimation de β sera vraisemblablement peu élevée par rapport à la variance due à l'imputation étant donné β (Rubin et Schenker 1986).

Bien qu'il soit possible en principe d'évaluer la variance due à l'imputation de codes d'appariement et de dénombrement au moyen de la méthode du TOR, on ne peut encore évaluer la variance due à l'imputation de caractéristiques manquantes (section 4). En fait, il serait possible d'évaluer cette variabilité en multipliant les caractéristiques d'imputation dans les échantillons P et D (Rubin 1987). Il faudrait alors calculer plusieurs estimateurs bivariés – un pour chaque série d'imputations.

Les modèles sur lesquels reposent les méthodes de pondération et d'imputation utilisées dans le TOR supposent que la probabilité qu'une variable soit manquante ne dépend pas de la valeur de celle-ci, étant donné les données observées. Une autre question ayant trait à l'imputation est de savoir quelle est la meilleure façon d'imputer simultanément des caractéristiques et des codes d'appariement (ou de dénombrement). La méthode utilisée dans le TOR (qui consiste à imputer tout d'abord les caractéristiques et ensuite les codes en fonction des caractéristiques imputées) suppose que les codes ne sont pas des variables explicatives utiles pour l'imputation de caractéristiques. Il serait peut-être plus indiqué d'adopter des modèles qui assouplissent les hypothèses du TOR. Rubin, Schaffer et Schenker (1988) approfondissent la question.

Les données manquantes constituent une seule source d'erreur dans l'estimation de l'erreur de couverture. D'autres sources, comme l'erreur d'appariement et le non-respect de l'hypothèse des probabilités de saisie constantes (section 2), sont analysées dans Hogan et Wolter (1988). Après avoir évalué toutes ces sources d'erreur pour le TOR, Hogan et Wolter concluent à la supériorité du TOR par rapport au dénombrement initial.

Tableau 3
Estimations des paramètres du modèle additif (2) servant à prévoir les taux de sous-dénombrement estimés du tableau 2

	Origine hispanique	Origine asiatique	Autre
α_0	9.82	7.31	6.21
α_p	-0.28	-0.7925	0.03
α_m	-0.505	-0.0675	-0.02
α_w	-0.525	-0.5525	-0.38

interposée, aux cas de déménagement et aux cas portant le code *W1* respectivement. Lorsqu'on utilise l'équation (2) pour prévoir les valeurs du tableau 2, le résiduel le plus élevé est 0.02%.

7.2 Méthode qui accroît le taux de sous-dénombrement estimé

Comme le TOR était limité à une très petite région des États-Unis, l'enquête postcensitaire n'a pas permis de recueillir des données sur les personnes qui avaient quitté la région d'essai après le recensement. Le fait d'exclure ces personnes de l'estimation revenait à supposer qu'elles avaient le même taux de saisie dans le recensement que les personnes incluses dans l'estimation. C'était là une hypothèse modérée puisque l'on sait que les personnes ayant déménagé ont habituellement un taux de saisie moindre que les personnes n'ayant pas déménagé.

Quatre cent neuf personnes sont venues s'établir dans la région d'essai entre le jour du recensement et le jour de l'EP. Ces nouveaux venus n'ont pas été inclus dans l'estimation puisque le jour du recensement, ils demeuraient à l'extérieur de la région d'essai et que, par conséquent, les données recueillies ce jour-là s'appliquent à d'autres régions. De plus, il n'était pas possible d'appartier ces nouveaux venus à des enrégistrement du recensement puisque ils vivaient à l'extérieur de la région d'essai le jour du recensement.

Pour avoir une idée des conséquences qu'aurait l'inclusion des sortants dans l'estimation, on pourrait considérer à la place les 409 entrants et imputer des probabilités d'appartenance pour ces personnes (puisque leurs codes d'appartenance sont inconnus). Les combinaisons qui ont produit les estimations les plus élevées et les plus faibles dans le tableau 2 ont été appliquées aux données du TOR en tenant compte des entrants; les résultats pertinents figurent dans le tableau 4. On remarquera que les estimations obtenues par l'application de méthodes autres

Tableau 4

Taux de sous-dénombrement estimés (en %) selon l'origine ethnique lorsque les entrants sont considérés dans l'estimation avec probabilités d'appartenance imputées

Indicateur de méthode (1 = autre, 0 = TOR)				
Interview par personne interposée	Personne ayant déménagé	<i>W1</i>	Origine hispanique	Origine asiatique
0	0	0	10.16	7.81
1	1	1	8.50	5.86
				5.81

NOTA: Les indicateurs des strates d'échantillonnage ont servi de variables explicatives dans les régressions logistiques pour l'imputation des probabilités d'appartenance et d'enrégistrement erroné.

Tableau 2

Taux de sous-dénombrement estimés (en %) selon l'origine ethnique et suivant diverses méthodes de traitement des cas d'interview par personne interposée de l'échantillon P, des cas de déménagement de l'échantillon P et des cas de l'échantillon D portant le code W1

Indicateur de méthode (1 = autre, 0 = TOR)				
Interview par personne interposée	Personne ayant déménagé	W1	Origine hispanique	Origine asiatique
Autre				
0	0	0	9.82	7.31
0	0	1	9.30	6.76
0	1	0	9.33	7.24
0	1	1	8.80	6.69
1	0	0	9.55	6.52
1	0	1	9.03	5.96
1	1	0	9.04	6.45
1	1	1	8.51	5.90
5.84				

NOTA: Les indicateurs des strates d'échantillonnage ont servi de variables explicatives dans les régressions logistiques pour l'imputation des probabilités d'appariement et d'enregistrement erroné.

tous les enregistrements portant le code W1 ont été vérifiées par des spécialistes de l'appariement. Ceux-ci identifiaient les enregistrements qui pouvaient laisser croire par un signe quelconque (par exemple: une note de l'interviewer) qu'il s'agissait d'un cas fictif; ils en ont relevé 118. Une méthode différente de celle utilisée dans le TOR consisterait à classer les 118 cas dans les enregistrements erronés résolus avant l'imputation. Cela aurait pour effet de hausser les taux observés et imputés d'enregistrement erroné.

Le tableau 2 donne les taux de sous-dénombrement estimés selon l'origine ethnique pour le plan factoriel 2x2x2, où les facteurs indiquent si des méthodes autres que celle utilisée dans le TOR ont été appliquées aux cas d'interview par personne interposée, aux cas de déménagement et aux cas portant le code W1. L'écart entre l'estimation la plus faible et l'estimation la plus élevée du taux de sous-dénombrement est de 1.31 point de pourcentage pour les personnes d'origine hispanique, de 1.41 point pour les personnes d'origine asiatique et de 0.43 point pour les autres.

Il convient de souligner qu'il y a peu d'interaction entre les méthodes de traitement des cas d'interview par personne interposée, des cas de déménagement et des cas portant le code W1 pour chaque origine ethnique. De fait, on peut utiliser le modèle additif simple ci-dessous pour prévoir les valeurs du tableau 2 pour chaque origine ethnique:

$$Y = \alpha_0 + I_p \alpha_p + I_m \alpha_m + I_w \alpha_w, \tag{2}$$

où Y est l'estimation prévue du taux de sous-dénombrement, I_p , I_m , et I_w sont les indicateurs de méthode (1 = autre, 0 = TOR) pour les cas d'interview par personne interposée, les cas de déménagement et les cas portant le code W1 respectivement et α_0 , α_p , α_m , et α_w sont les estimations de paramètres données dans le tableau 3. Le paramètre α_0 est le taux de sous-dénombrement estimé lorsqu'aucune autre méthode que celle du TOR n'est utilisée; α_p , α_m , et α_w sont les effets de l'application d'autres méthodes aux cas d'interview par personne

7. ESTIMATION DE L'ERREUR DE COUVERTURE SUIVANT D'AUTRES MÉTHODES DE TRAITEMENT DES DONNÉES MANQUANTES ET AUTRES SOURCES DE PROBLÈME

Dans cette section, nous analysons l'incidence d'autres méthodes de traitement des données manquantes et autres sources de problème sur les estimations de l'erreur de couverture pour les trois catégories définies par la variable Origine Ethnique (hispanique, asiatique et autre). Pour une méthode de traitement et une catégorie données, soient N la somme des estimations bivariées pour toutes les strates formées à posteriori qui correspondent à la catégorie donnée et N_c la somme des chiffres de recensement non rajustés pour ces strates. Le taux de sous-dénombrement estimé est alors $100(1 - N_c / N)\%$.

Nous envisageons tout d'abord d'utiliser les indicateurs des strates d'échantillonnage comme variables explicatives dans les régressions logistiques appliquées aux échantillons P et D pour l'imputation des probabilités d'appariement et d'enregistrement erroné traitées dans les sections 5 et 6. Le TOR a produit des estimations de taux de sous-dénombrement sans faire intervenir ces variables explicatives; ces taux sont 9.85% pour les personnes d'origine hispanique, 7.32% pour les personnes d'origine asiatique et 6.24% pour les autres. Lorsqu'on utilise les indicateurs des strates d'échantillonnage, les taux deviennent 9.82% pour les personnes d'origine hispanique, 7.31% pour les personnes d'origine asiatique et 6.21% pour les autres. L'utilisation des indicateurs de strates d'échantillonnage ne modifie les taux initiaux que de 0.03 point de pourcentage au maximum. Néanmoins, cette utilisation sera considérée dans toutes les méthodes que nous analyserons car elle produit en principe des résultats plus précis; par exemple, on devrait obtenir des erreurs types plus justes.

7.1 Méthodes qui réduisent le taux de sous-dénombrement estimé

Le taux d'appariement pour les 375 cas d'interview par personne interposée de l'échantillon P qui ont été résolus était de 78.9% comparativement à 87.8% pour l'ensemble de l'échantillon P . Même s'il est probable que les cas d'interview par personne interposée aient eu un taux de saisie moins élevé que les cas d'interview ordinaire dans le recensement, l'écart entre les taux d'appariement est peut-être attribuable en partie à l'absence ou à l'inexactitude de données provenant d'interviews par personne interposée (voir section 4). Une façon prudente de procéder consisterait à classer les 189 interviews par personne interposée dans les non-interviews et à appliquer la méthode de pondération décrite dans la section 3; ainsi, les cas d'interview par personne interposée auraient essentiellement le même taux d'appariement que les cas d'interview ordinaire. (Notons que lorsque toutes les interviews par personne interposée sont considérées comme des non-interviews, le code d'interview ne figure plus dans le modèle de régression logistique pour l'imputation des probabilités d'appariement.)

Le taux d'appariement pour les 277 cas de déménagement (entre le jour du recensement et le jour de l'EP) de l'échantillon P qui ont été résolus était de 66.1%. Bien que l'on croit généralement que les personnes ayant déménagé ont un taux de saisie moins élevé dans le recensement que les personnes n'ayant pas déménagé, le faible taux d'appariement observé dans le cas des personnes ayant déménagé est peut-être attribuable en partie aux difficultés inhérentes à l'appariement de cette catégorie de personnes, par exemple la difficulté d'obtenir l'adresse à laquelle demeurait la personne le jour du recensement. Une façon prudente de procéder consisterait à classer tous les cas de déménagement dans les cas non résolus puis à imputer des probabilités d'appariement pour les cas non résolus à l'aide d'un modèle de régression logistique qui n'utilise pas le code "déménagement" comme variable explicative. Ainsi, les personnes ayant déménagé auraient essentiellement le même taux d'appariement que les personnes n'ayant pas déménagé.

Sur les 979 cas non résolus de l'échantillon D , 257 portaient le code d'interview de suivi $W1$, qui signifiait que le répondant ne connaissait pas la personne en question. Ce code pouvait aussi indiquer que la personne en question était un individu fictif. C'est pourquoi, après le TOR,

indiquent que les probabilités d'appariement imputées sont beaucoup moins élevées dans le cas des interviews par personne interposée et des personnes n'ayant pas déménagé. Cette situation s'explique peut-être en partie par la difficulté d'apparier les cas d'interview par personne interposée et les cas de déménagement plutôt que par le seul fait que les taux de saisie du recensement sont moins élevés dans ces cas. Si cela est exact, le temps est peut-être venu de considérer d'autres méthodes de traitement des données manquantes; c'est ce à quoi nous nous attachons dans la section 7.

Sur les 19,391 cas résolus de l'échantillon P , 17,018 (87.8%) se sont avérés des cas de concordance. La somme (non pondérée) des 161 probabilités d'appariement imputées était de 124.66 (pour un taux d'appariement imputé de 77.4%). Même si l'on a fait usage d'un échantillon stratifié d'échantillons dans le TOR, l'estimation des paramètres de régression logistique a supposé un échantillon aléatoire simple de personnes. Afin de vérifier si l'omission de la stratification n'aurait pas créé de biais, nous avons ajusté une fois de plus le modèle de régression logistique (une fois le TOR achevé) en incluant dans X les variables indicatrices des six strates d'échantillonnage (Diffendal 1988). Nous avons refait la somme des probabilités d'appariement imputées et nous avons obtenu 124.50 (77.3%). La faible incidence de ce changement sur les estimations de l'erreur de couverture dans le recensement est démontrée dans la section 7. Dans la section 8, nous examinons les conséquences des effets de plan qui pourraient découler de la formation de grappes.

6. CODES DE DÉNOMBREMENT MANQUANTS DANS L'ÉCHANTILLON D

Sur les 20,976 cas de l'échantillon D , 3,714 ont fait ou auraient dû faire l'objet d'un suivi. Après le suivi, 979 cas (4.7% de l'effectif, 26.4% des cas devant faire l'objet d'un suivi) n'avaient toujours pas de code de dénombrement. Tous ces cas non résolus sauf neuf pouvaient être classés dans quatre grandes catégories: d'abord, 498 cas qui n'ont pas fait l'objet d'un suivi alors qu'ils auraient dû; ensuite, 257 cas où la personne qui a participé à l'interview de rappel ne connaissait pas la personne en question, puis 137 cas pour lesquels l'interview n'a pas produit assez de renseignements pour déterminer un code de dénombrement, et enfin 78 cas pour lesquels il n'a pas été possible d'obtenir une interview de rappel.

On a pallié à l'absence de codes de dénombrement dans l'échantillon D en imputant une probabilité d'enregistrement erroné pour chaque cas non résolu. Dans le terme EB de l'expression définissant l'estimateur bivalent (équation 1), les cas non résolus sont représentés par la somme pondérée des probabilités imputées. La méthode d'imputation différerait peu de celle utilisée pour les codes d'appariement. Un seul changement majeur: comme l'absence de codes de dénombrement ne pouvait être constatée qu'à la suite des procédures de suivi, seuls les cas résolus par suivi ont servi à l'estimation du modèle de régression logistique. Les variables de base qui ont servi à définir X pour la régression logistique sont le Mode d'Occupation, le Type de Construction, le Sexe, l'Âge et l'Origine Ethnique, de même que les variables qui indiquent si le questionnaire du recensement concernant le ménage dont fait partie la personne visée a été retourné par la poste ou si le ménage au complet ou une partie de celui-ci n'a pu être apparié avant le suivi. Le tableau 2 (dans l'annexe) donne les estimations des coefficients de la régression logistique.

Sur les 17,262 cas qui ne nécessitaient pas de suivi, 278 (1.6%) ont été classés parmi les enregistrements erronés ou les cas non appariables. Deux mille sept cent trente-cinq cas ont été résolus par suivi et 82 de ces cas (3.0%) ont été classés dans les enregistrements erronés. La somme (non pondérée) des 979 probabilités imputées était de 21.93 (2.2%). Lorsqu'on inclut les variables indicatrices des strates d'échantillonnage dans X , la somme des probabilités imputées devient 23.58 (2.4%). Comme dans le cas de l'échantillon P , ce changement a une très faible incidence sur les estimations de l'erreur de couverture; voir section 7.

5. CODES D'APPARIEMENT MANQUANTS DANS L'ÉCHANTILLON P

Sur les 19,552 enregistrés de l'échantillon P créés à la suite d'une interview complète, 161 (0,8%) n'avaient pas de code d'appariement pour l'estimation "bivalente". Tous ces cas non résolus saut trois pouvaient être classés dans deux grandes catégories: d'abord, 105 ménages pour lesquels on n'a pas tenté d'appariement à cause de noms incomplets ou de caractéristiques insuffisantes, puis 53 autres ménages qui ont déménagé entre le jour du recensement et le jour de l'EP et pour lesquels il a été difficile de déterminer l'adresse le jour du recensement ou de retracer le questionnaire de recensement.

Une façon classique de pallier l'absence d'un élément binaire comme le code d'appariement est d'imputer l'une des deux valeurs possibles. Par exemple, lorsqu'on estime le taux de sous-dénombrement pour le recensement décennal de 1980, on a imputé un code d'appariement pour chaque cas de l'échantillon P non résolu en se fondant sur un cas résolu qui présentait les mêmes caractéristiques (Fay, Passel et Robinson 1988, Chapitre 6). Toutefois, une méthode difficile- rente a été appliquée dans le TOR. Après avoir imputé toutes les caractéristiques manquantes à l'aide des méthodes décrites dans la section 4, on a imputé une probabilité d'appariement pour chaque code manquant, la probabilité étant estimée au moyen d'un modèle explicite qui sera décrit un peu plus loin. Dans le dénominateur de l'expression qui définit l'estimateur biva- lent (équation 1), les cas non résolus sont représentés par la somme pondérée des probabilités imputées.

On a imputé des probabilités plutôt que des valeurs binaires pour deux raisons. La première est que l'imputation de valeurs binaires aléatoires est moins efficace que l'imputation de pro- babilités estimées puisqu'elle produit des estimations qui ont une variance plus élevée (voir Rubin 1987, p. 15). En deuxième lieu, comme les probabilités imputées représentent une incer- titude à propos des codes d'appariement manquants, ces probabilités devraient pouvoir servir à calculer une variance due à l'imputation. Cependant, comme l'estimateur bivalent (1) est non linéaire en M, l'imputation d'une probabilité (ou d'une moyenne) pour chaque élément binaire manquant introduit un biais dans l'estimation (voir Rubin 1987, p. 14). Les statisti- ciens approfondissent actuellement l'utilisation de probabilités imputées pour des données binaires manquantes.

La méthode de régression logistique suivante a servi à imputer les probabilités d'apparie- ment. Soit X un vecteur de variables explicatives, Y = concordance ou non-concordance, et $p = \text{Pr}(Y = \text{concordance} | X)$. Le vecteur de paramètres β du modèle de régression logistique

$$\text{logit}(p) = \log[p / (1 - p)] = X' \beta$$

a été estimé sur la base des données relatives aux cas résolus à l'aide des méthodes bayésiennes pour régression logistique catégorique décrites dans Rubin et Schenker (1987); ces méthodes consistent à additionner des observations fractionnaires à chaque case ou cellule dans la régres- sion logistique puis à ajuster le modèle par les méthodes classiques du maximum de vraisem- blance. Ainsi, pour le cas non résolu j, étant donné $X = x_j$, la probabilité d'appariement imputée a été définie par

$$\hat{p}_j = \text{logit}^{-1}(x_j' \beta) = \exp(x_j' \beta) / [1 + \exp(x_j' \beta)] ,$$

où β désigne l'estimateur de β . Les variables de base qui ont servi à définir X sont le Mode d'Occupation, le Type de Construction, le Sexe, l'Âge et l'Origine Ethnique, de même que les variables qui indiquent s'il s'agit d'une interview ordinaire ou d'une interview par personne interposée ou s'il s'agit d'une personne qui a déménagé ou d'une personne qui n'a pas démé- nagé entre le jour du recensement et le jour de l'EP.

Le tableau A1 (dans l'annexe) donne les estimations des coefficients de la régression logis- tique. Les coefficients élevés qui se rattachent au code d'interview et au code "déménagement"

4. CARACTÉRISTIQUES MANQUANTES DANS LES ÉCHANTILLONS P ET D

Même lorsqu'on avait pu réaliser une interview avec un ménage de l'échantillon P, les données concernant les caractéristiques des personnes et du logement étaient parfois incomplètes. Il y avait aussi des données incomplètes dans le recensement et, partant, dans l'échantillon D. Les variables utilisées pour la stratification a posteriori dans le TOR (Diffendal 1988) comprenaient la variable de logement Mode d'Occupation (1 = logement occupé par le propriétaire, 2 = logement loué ou occupé sans contrepartie) et les variables personnelles Sexe (1 = homme, 2 = femme), Âge (1 = 0-14, 2 = 15-29, 3 = 30-44, 4 = 45-64, 5 = 65 +) et Origine Ethnique (1 = origine hispanique, 2 = origine asiatique, 3 = autre). En outre, la variable de logement Type de Construction (1 = maison unifamiliale, 2 = immeuble à logements multiples) a servi à traiter les codes d'appariement manquants dans l'échantillon P et les codes de dénombrement manquants dans l'échantillon D (voir sections 5 et 6).

Le tableau 1 renferme les chiffres concernant les caractéristiques manquantes pour les échantillons P et D au complet et pour les cas d'interview par personne interposée de l'échantillon P. En ce qui concerne les échantillons P et D, le taux de données manquantes le plus élevé est de 7,0% (pour la variable Origine Ethnique dans l'échantillon D), tous les autres taux n'excédant pas 3,5%. Les taux de données manquantes pour les interviews par personne interposée de l'échantillon P sont tous plusieurs fois supérieurs aux taux observés pour l'échantillon P en général, quoique seule la variable Mode d'Occupation (20,2%) présente un taux de données manquantes supérieur à 10%.

Les caractéristiques manquantes pour chacun des échantillons (P et D) ont été imputées par une méthode "hot-deck" comportant deux lectures de données après que les données ont été classées selon des critères géographiques. À la première lecture, les valeurs manquantes des variables Mode d'Occupation, Type de Construction et Origine Ethnique ont été imputées à l'aide des observations les plus récentes à cause de la relation étroite qui est censée exister entre ces variables et les caractéristiques géographiques. En outre, on a établi la distribution des variables Sexe et Âge en fonction du genre de ménage (ménage à une personne ou à plusieurs personnes), de l'état matrimonial, du lien avec le chef de ménage et du sexe et de l'âge du chef de ménage en se servant de toutes les observations. À la seconde lecture, les valeurs manquantes des variables Sexe et Âge ont été imputées aléatoirement à partir des distributions établies au cours de la première lecture. Pour plus de détails sur l'imputation des caractéristiques dans le TOR, voir Schenker (1987).

En résumé, le plan de sondage de l'EP (fondée sur un échantillon d'îlots) a servi non seulement à l'élaboration d'un schéma de pondération pour non-interviews (section 3) mais aussi à l'imputation de caractéristiques qui tendent à être réunies dans un même îlot, c'est-à-dire Mode d'Occupation, Type de Construction et Origine Ethnique.

Tableau 1

Chiffres sur les caractéristiques manquantes
(pourcentages entre parenthèses) pour les échantillons P et D
et les cas d'interview par personne interposée de l'échantillon P

Variable	Echantillon P (19,552 personnes)	Echantillon D (20,976 personnes)	Interviews par personne interposée de l'échantillon P (430 personnes)
Mode d'Occupation	690 (3.5)	154 (0.7)	87 (20.2)
Type de Construction	459 (2.3)	343 (1.6)	38 (8.8)
Sexe	418 (2.1)	82 (0.4)	18 (4.2)
Âge	137 (0.7)	432 (2.1)	18 (4.2)
Origine Ethnique	155 (0.8)	1463 (7.0)	17 (4.0)

NOTA: Les 19,552 personnes de l'échantillon P comprennent les 430 cas d'interview par personne interposée.

où N_p est l'estimateur de la population de l'EP, CEN est le chiffre du recensement non rajusté, SUB est le nombre de substitutions de personnes (pour les cas de non-réponse au questionnaire) dans le recensement, BE est un estimateur du nombre d'enregistrements erronés et de personnes non attachées dans le recensement et M est l'estimateur du nombre de personnes qui font partie à la fois de la population du recensement et de la population de l'EP, les valeurs CEN et SUB sont déduites des données du recensement tandis que N_p , BE et M sont établis à partir des données des échantillons P et D . De fait, le calcul de l'estimateur bivalent équivalent à multiplier le nombre estimé d'enregistrements exacts et attachés du recensement (CEN-SUB-BE) par l'inverse du taux de saisie estimé du recensement (M/N_p). La théorie de l'estimation "bivalente" suppose que la probabilité de saisie est la même pour tous les membres du domaine auquel s'applique l'estimateur et ce, aussi bien pour le recensement que pour l'enquête postcensitaire (Wolter 1986). Par conséquent, un groupe quelconque de personnes incluses dans le domaine n'aura pas plus de chances ni moins de chances que n'importe quel autre groupe d'être dénombré dans le recensement ou l'EP. Pour rendre cette hypothèse plus réaliste dans le TOR, on a calculé des estimateurs bivalents à l'intérieur de strates formées à posteriori en fonction des caractéristiques des personnes et du logement. Les strates formées à posteriori sont décrites dans Diffendal (1988). Il peut s'agir par exemple des locataires de sexe masculin et d'origine hispanique qui sont âgés de 30 à 44 ans et vivent dans des ilots à majorité hispanique.

En résumé, les données des échantillons P et D qu'il est nécessaire de connaître pour estimer l'erreur de couverture sont le code d'appariement (concordance vs. non-concordance) de chaque membre de l'échantillon P , le code de dénombrement (exact vs. erroné) de chaque membre de l'échantillon D et les caractéristiques des personnes et du logement pour chacun des membres des deux échantillons.

3. MÉNAGES DE L'ÉCHANTILLON P NON INTERVIEWÉS

Il se peut qu'un intervieweur n'ait pas été en mesure de réaliser une interview pour une unité de logement occupée à cause, par exemple, d'un refus de répondre de la part des occupants. Sur les 5,935 unités de logement qui étaient considérées comme occupées, 32 (0.5%) ont été classées parmi les cas de non-interview. À cause de ces cas de non-interview, il n'a pas été possible d'obtenir toutes les données voulues concernant le nombre de personnes dans chaque ménage, les caractéristiques des personnes et du logement et les codes d'appariement. Grâce au plan de sondage de l'EP (fondée sur un échantillon d'ilots), il a été facile de résoudre le problème de la non-interview de ménages de l'échantillon P . Dans chaque ilot échantillonné, les poids de sondage des ménages de non-interview ont été redistribués parmi les ménages interviewés. La répartition pour non-interview suppose essentiellement que la distribution des personnes, des caractéristiques et des codes d'appariement pour les ménages non interviewés à l'intérieur d'un ilot est la même que pour les ménages interviewés. On a utilisé cette hypothèse car les ménages d'un même ilot tendent à avoir plus de caractéristiques en commun que les ménages d'ilots différents, bien qu'il existe probablement des différences notables entre les ménages de non-interview et les ménages interviewés, surtout au point de vue de la taille (voir, par ex., Palmer 1967).

Il se peut que les données recueillies sur un ménage au moyen d'une interview par personne interviewée (dans le TOR, cela signifie une interview complète réalisée avec une personne qui n'est pas membre du ménage) soient d'une qualité tellement faible que l'on est porté à classer ce ménage dans les ménages de non-interview. La qualité des données provenant des 189 interviews par personne interviewée du TOR est analysée dans la section 4 tandis que la section 7 renferme quelques estimations de l'erreur de couverture établies lorsque les interviews par personne interviewée sont considérées comme des non-interviews.

5,86 et 7,81% pour les personnes d'origine asiatique et 5,81 et 6,59% pour les autres personnes. Les estimations correspondantes calculées à l'aide du TOR étaient 9,85, 7,32 et 6,21% respectivement. Enfin, la section 8 présente une analyse et les conclusions.

2. ESTIMATION DE L'ERREUR DE COUVERTURE DANS LE RECENSEMENT

Diffendal (1988) expose en détail la méthode d'estimation de l'erreur de couverture du recensement utilisée dans le TOR. La présente section expose brièvement les points qu'il est nécessaire de connaître pour comprendre le reste de cet article. L'erreur de couverture a été estimée à l'aide de données d'une enquête postcensitaire (EP) menée auprès de personnes demeurant dans la région de recensement. On a commencé par prélever un échantillon d'îlots dans la région. Ensuite, chaque unité de logement contenue dans ces îlots a fait l'objet d'une enquête afin d'en connaître les occupants le jour du recensement, les occupants le jour de l'EP (et l'endroit où ces derniers demeuraient le jour du recensement) et de déterminer les caractéristiques de ces occupants.

Deux échantillons ont servi à estimer l'erreur de couverture du recensement. L'échantillon P (pour population) était composé des personnes qui demeuraient dans les îlots échantillonnés de l'EP le jour de cette enquête. On tentait d'appartier chaque membre de l'échantillon P à une personne qui avait été recensée afin de déterminer si ce membre avait lui-même été recensé; le taux d'appariement à l'intérieur de chaque domaine d'analyse servait essentiellement à estimer le taux de saisie du recensement pour ce domaine. L'échantillon D (pour dénombrement) était composé des personnes qui demeuraient dans les îlots échantillonnés de l'EP le jour du recensement; cet échantillon a servi à estimer le nombre d'enregistrements erronés (par ex., enregistrements fictifs ou redoublés) et de personnes non appariables (par ex., personnes pour lesquelles aucun nom n'était indiqué) dans le recensement à l'intérieur de chaque domaine. On tentait d'appartier chaque membre de l'échantillon D à une personne qui avait participé à l'enquête postcensitaire. Chaque cas d'appariement était considéré comme un enregistrement exact puisque l'EP indiquait que la personne devait avoir été recensée. Chaque cas de non-appariement était approfondi pour déterminer s'il s'agissait d'un enregistrement erroné ou d'un enregistrement exact qui avait été oublié dans l'EP (laquelle n'est pas censée être à l'abri des erreurs de couverture).

Si une EP était réalisée à la grandeur des États-Unis, les membres de l'échantillon P qui ont quitté le Central Los Angeles County entre le jour du recensement et le jour de l'EP participeraient au recensement et le jour de l'EP participeraient au recensement. On tenterait ensuite d'appartier ces personnes aux enregistrements du Central Los Angeles County faits le jour du recensement et on se servirait des résultats pour estimer l'erreur de couverture dans cette région. De même, les membres de l'échantillon P qui se sont établis dans le Central Los Angeles County entre le jour du recensement et le jour de l'EP contribueraient à l'estimation de l'erreur de couverture à l'extérieur de cette région. Toutefois, comme le recensement et l'EP pour le TOR ne s'appliquaient qu'au Central Los Angeles County (et non à tous les États-Unis), les personnes qui avaient quitté la région d'essai ne participaient pas à l'EP et celles qui s'y étaient établies n'étaient pas concernées. Par conséquent, les données relatives à ces deux catégories de personnes n'ont pas servi à l'estimation. (Notons toutefois que les données relatives aux personnes qui ont déménagé sans quitter la région d'essai ont servi à l'estimation). Cette question est analysée plus à fond dans la section 7.2.

L'estimateur "bivalent" de la taille de la population (voir Marks, Selzer et Krotki 1974; Krotki 1978 et Wolter 1986 pour analyse et bibliographie) est défini par

$$DSE = N_p (CEN-SUB-EE) / M,$$

(1)

Traitement des données manquantes dans l'estimation de la couverture: le test des opérations de redressement de 1986

NATHANIEL SCHENKER¹

RÉSUMÉ

Cet article porte sur les méthodes de traitement des données manquantes dans les enquêtes postcensitaires en vue de l'estimation de l'erreur de couverture dans le recensement; à titre d'illustration, nous analysons le test des opérations de redressement de 1986 (Diffendal 1988). Les méthodes précitées comprennent des méthodes d'imputation fondées sur le hot-deck et des modèles de régression logistique de même que des méthodes de redressement par la pondération. Nous analysons également la sensibilité des estimations de sous-dénombrement tirées du test de 1986 à la variation des modèles d'imputation.

MOTS CLÉS: Imputation; non-réponse; enquête postcensitaire; redressement par pondération; sous-dénombrement.

1. INTRODUCTION

Les données manquantes peuvent être une importante source d'incertitude dans l'estimation de l'erreur de couverture pour les recensements décennaux aux États-Unis (Freedman et Navidi 1986; Fay, Passel et Robinson 1988, Chapitre 6). Pour les recensements décennaux de 1960 et de 1980, diverses méthodes de traitement des données manquantes ont produit plusieurs estimations de l'erreur de couverture.

Le Census Bureau a appliqué de nombreux tests à des méthodes d'estimation de l'erreur de couverture pour être en mesure de traiter le problème des données manquantes et les autres problèmes qui pourraient surgir au cours du recensement décennal de 1990. Un de ces tests a été le Test des opérations de redressement (TOR) de 1986 (Diffendal 1988), qui était fondé sur le recensement de 1986 du Central Los Angeles County. Une modification des méthodes appliquées sur le terrain et du plan de sondage du TOR a permis de réduire la quantité de données manquantes par rapport à 1980 (Hogan et Wolter 1988). Malgré cela, certains problèmes persistent.

Cet article vise à décrire les problèmes qu'ont posés les données manquantes dans le TOR et la façon dont ils ont été traités dans le processus d'estimation. Dans la section 2, nous décrivons brièvement la méthode d'estimation de l'erreur de couverture appliquée dans le TOR. Dans les sections 3 à 6, nous analysons le genre et la quantité de données manquantes de même que les méthodes utilisées pour résoudre ce problème, notamment une méthode de redressement par pondération pour les cas de non-réponse au questionnaire (non-interview), une méthode d'imputation "hot-deck" pour les caractéristiques démographiques et les caractéristiques du logement manquantes ainsi qu'une méthode d'imputation fondée sur des modèles de régression logistique pour certains éléments binaires ayant trait au dénombrement dans le recensement. La section 7 présente des estimations de l'erreur de couverture calculées selon divers modèles d'imputation et diverses méthodes de traitement appliquées à des cas précis. Les estimations minimum et maximum du taux de sous-dénombrement calculées à l'aide de ces modèles ou de ces méthodes sont 8.50 et 10.16% pour les personnes d'origine hispanique,

¹ Nathaniel Schenker, Undercount Research Staff, Statistical Research Division, Bureau of the Census, Washington (DC) 20233, E.-U. Cet article est un compte rendu des recherches qui ont été faites par un membre du personnel du Census Bureau. Les opinions qui y sont exprimées sont celles de l'auteur et ne reflètent pas nécessairement la position du Census Bureau.

- Diffendal: Test des opérations de redressement
- MULRY, M.H., HOGAN, H.R., WALKER, J.R., CHAPMAN, D.R., EVAUL, J., et MOORE, R.H. (1981). Research proposal for a study of methods for 1990 decennial census coverage evaluation. Report technique, U.S., Bureau of the Census, Washington, D.C.
- SCHENKER, NATHANIEL (1988). Traitement des données manquantes dans l'estimation de la couverture: le test des opérations de redressement de 1986. *Techniques d'enquête*, 14.
- U.S. Bureau of the Census (1979). Evaluation. Dans *POSTAN: Program Considerations*, Partie A, Chapitre A-13.
- WOLFGANG, GLENN (1987). The pre-enumeration survey of the 1986 census of Central Los Angeles County. Article présenté à la conférence annuelle de l'American Statistical Association, San Francisco.
- WOLTER, KIRK M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, 81, 338-346.
- ZASLAVASKY, ALAN M. (1988). Representing local area undercount by reweighing of households. Proceedings of the Bureau of the Census Fourth Annual Research Conference.

sexe. Nous avons lissé ces estimations en leur ajustant un modèle de régression puis nous avons reproduit les estimations ainsi obtenues au niveau de l'ilot. L'utilisation d'un taux de sous-dénombrement estimé pour l'ilot permet l'agrégation de données à n'importe quel niveau au-dessus de l'ilot.

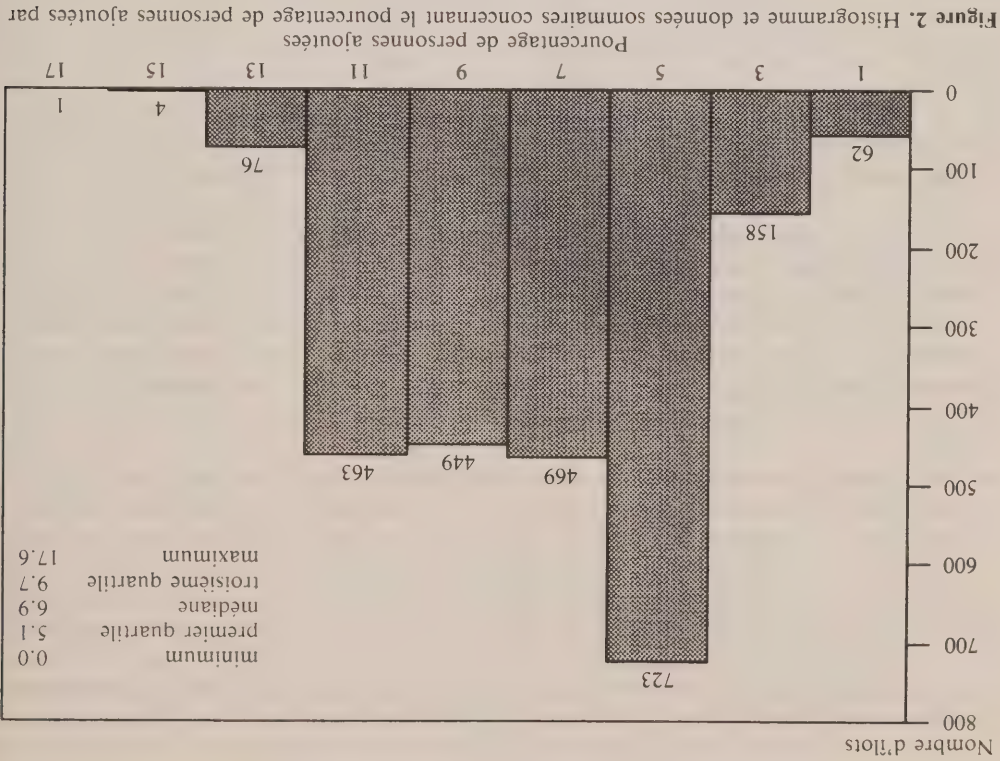
Schenker (1988) et Hogan et Wolter (1988) font une évaluation des opérations et définissent les hypothèses sur lesquelles reposent les estimateurs. Ces deux ouvrages ainsi que le précédent article constituent une analyse approfondie des chiffres du recensement et des taux de sous-dénombrement estimés pour le recensement d'essai.

REMERCIEMENTS

L'auteur tient à exprimer sa reconnaissance aux personnes qui ont collaboré à cette étude: Dan Childers, Howard Hogan, Cary Isaki, Matthew Jaro, Jan Jaworski, Charisse Jeffries, Robert O'Brien, Arona Pistiner, Nathaniel Schenker, Maria Urrutia, Debbie Wagner et Kirk Wolter. L'auteur tient aussi à remercier l'arbitre, dont les commentaires ont permis d'améliorer le contenu de cet article.

BIBLIOGRAPHIE

- ANOLIK, IRWIN (1988). The 1986 rural post-enumeration survey in East Central Mississippi. Statistical Research Division Report Series. CENSUS/SRD/RR-87/09.
- CAUSEY, B.D., COX, L.H., et ERNST, L.R. (1985). Applications of transportation theory to statistical problems. *Journal of the American Statistical Association*, 80, 903-909.
- CITRO, CONSTANCE F., et COHEN, MICHAEL L. (1985). *The Bicentennial Census — New Directions for Methodology in 1990*. Washington: National Academy Press.
- ERICKSEN, EUGENE P., et KADANE, JOSEPH B. (1985). Estimating the population in a census year — 1980 and beyond. *Journal of the American Statistical Association*, 80, 98-109.
- FAY, R.E., PASSEL, J.S., et ROBINSON, J.G. (1988). The coverage of population in the 1980 census. 1980 Census of Population and Housing Evaluation and Research Report PHC80-E4, Washington: U.S. Government Printing Office.
- FREEDMAN, D. et NAVIDI, W. (1986). Models for adjusting the census. *Statistical Science*, 1, 3-11.
- HOGAN, HOWARD R. (1984). Research plan on adjustment. *Proceedings of the Social Statistics Section American Statistical Association*, 452-457.
- HOGAN, HOWARD R., et WOLTER, KIRK M. (1988). Mesure de l'erreur dans une enquête post-censitaire. *Techniques d'enquête*, 14.
- ISAKI, CARY T., SCHULTZ, LINDA K., SMITH, PHILIP J., et DIFFENDAL, GREGG J. (1987). Small area estimation research for census undercount-progress report. Dans *Small Area Statistics: An International Symposium*, 219-238. New York: John Wiley and Sons.
- JARO, MATTHEW, et CHILDERS, DANNY (1986). Matching the 1985 census of Tampa. Article présenté à la conférence annuelle de l'American Statistical Association. Chicago.
- MARKS, ELI S. (1978). The use of dual system estimation in census evaluation. Dans *Developments in Dual System Estimation*, (éd. Karol Krotki), Edmonton: University of Alberta Press.



Pour résumer les opérations de redressement au niveau de l'îlot, nous représentons dans la figure 1 le nombre de personnes qui ont été ajoutées selon le nombre d'îlots concernés. Près de 80% des îlots ont reçu moins de 20 personnes. Seulement deux îlots ont reçu plus de 150 personnes. Ces deux îlots étaient assez denses puisqu'ils comptaient chacun environ 2,000 personnes. Plus de 80% des îlots avaient un taux de sous-dénombrement estimé variant de 4% à 12%. Nombre de petits îlots ont reçu un faible pourcentage des personnes additionnelles parce que les estimations avaient été arrondies vers le bas et que cela s'était traduit par une forte variation du pourcentage. Les îlots qui ont reçu le plus fort pourcentage de personnes additionnelles étaient ceux qui comptaient surtout des locataires d'origine hispanique; du reste, ces îlots étaient ceux pour lesquels les facteurs de redressement prévus étaient les plus élevés.

5. CONCLUSION

Cet article nous a permis d'étudier les méthodes, les opérations et les résultats du Test des opérations de redressement (TOR). Ce test visait à vérifier le calendrier et les caractéristiques des opérations de redressement des données du recensement en regard à la population non dénombrée. Les résultats du TOR prouvent qu'il est possible de produire des estimations du taux de sous-dénombrement dans des délais raisonnables. En outre, le TOR a été réalisé plus rapidement que n'importe quelle enquête post-censitaire antérieure. Les résultats du TOR indiquent un taux de sous-dénombrement de 9% pour le recensement d'essai du Central Los Angeles County. Nous avons établi des estimations bivalentes pour 70 catégories formées en fonction de l'origine ethnique et du mode d'occupation, de l'âge et du

Le tableau 6 renferme les facteurs de redressement fondés sur l'échantillon (facteurs estimés) et les facteurs de redressement prévus pour les 70 strates formées a posteriori. En règle générale, les facteurs de redressement prévus sont inférieurs aux facteurs estimés les plus élevés mais supérieurs aux facteurs estimés les plus faibles. Les facteurs de redressement prévus ont une moins grande variabilité que les facteurs de redressement fondés sur l'échantillon. Les effets du modèle de régression sont les plus visibles dans le cas des locataires asiatiques de sexe féminin âgés de 65 et plus. Le facteur de redressement prévu pour cette strate est de 1.087 comparativement à un facteur de redressement estimé de 1.212. Dans ce cas, le facteur de redressement prévu est plus conforme à la représentation que l'on se fait du taux de sous-dénombrement pour les personnes de 65 ans et plus, c'est-à-dire un taux moins élevé que pour les autres groupes d'âge.

On a multiplié les chiffres du recensement par les facteurs de redressement prévus pour les 2 405 îlots de la région d'essai. On a ensuite arrondi les chiffres redressés pour obtenir des nombres entiers. Malgré que trois facteurs de redressement prévus aient été inférieurs à 1 (surdénombrement estimé), l'arrondissement n'a pas produit d'estimations gonflées. Les opérations de redressement ont permis d'ajouter 32,843 noms à la liste des personnes recensées (soit un taux de sous-dénombrement de 8.2%). Si l'on avait utilisé les facteurs de redressement fondés sur l'échantillon, c'est 36,454 noms que l'on aurait ajouté à la liste des personnes recensées (soit un taux de sous-dénombrement de 9.0%). Le lissage a donc réduit de près de 10% le niveau de sous-dénombrement estimé. Cette forte réduction s'explique par le fait que les groupes qui présentaient les estimations de sous-dénombrement les plus élevées avant le lissage étaient ceux qui comptaient le plus grand nombre de personnes.

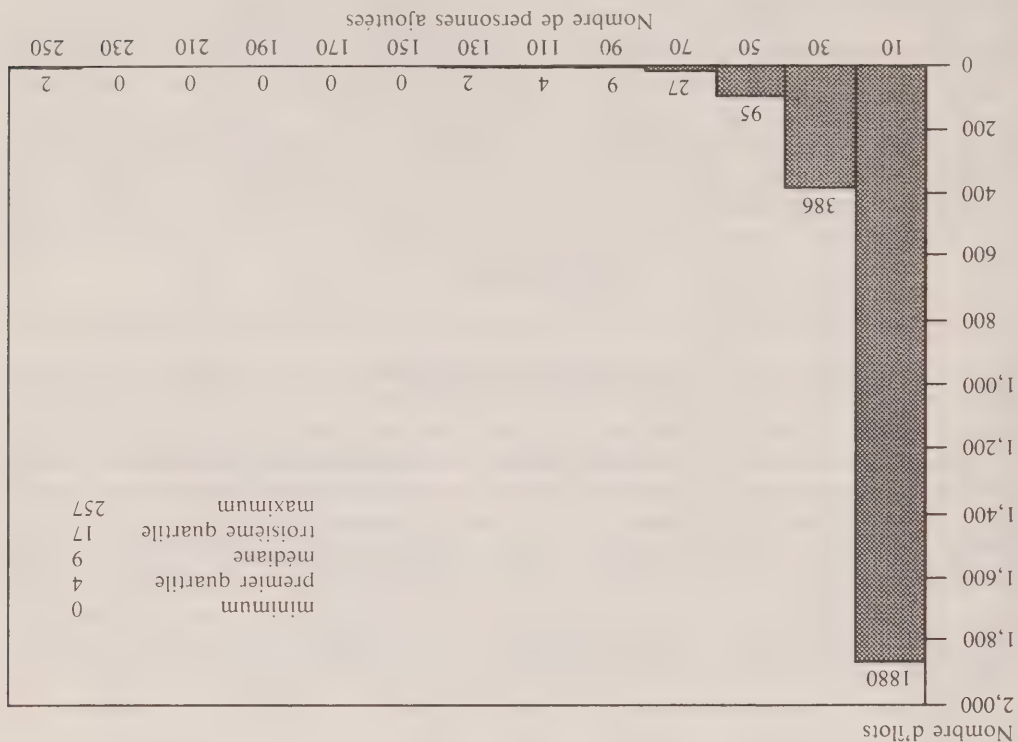


Figure 1. Histogramme et données sommaires concernant le nombre de personnes ajoutées par îlot

Tableau 6
Résultats du lissage des facteurs de redressement du TOR — Fin

Strates formées a posteriori	Sexe/Âge	Estimé		Prévu	
		Facteur de redressement (Y)	Erreur type	Facteur de redressement (4F)	Erreur type

PAS dans tous I	H' 0-14	1.045	0.030	1.041	0.019
PAS dans tous I	H' 15-29	1.059	0.038	1.085	0.022
PAS dans tous I	H' 30-44	1.091	0.040	1.053	0.022
PAS dans tous I	H' 45-64	1.035	0.020	1.033	0.016
PAS dans tous I	H' 65 +	1.031	0.051	1.037	0.023
PAS dans tous I	F 0-14	1.040	0.041	1.039	0.022
PAS dans tous I	F 15-29	1.052	0.046	1.086	0.024
PAS dans tous I	F 30-44	1.035	0.036	1.037	0.021
PAS dans tous I	F 45-64	1.038	0.019	1.035	0.015
PAS dans tous I	F 65 +	1.051	0.045	1.041	0.022
AL dans tous I	H' 0-14	1.037	0.059	1.049	0.027
AL dans tous I	H' 15-29	1.252	0.114	1.115	0.031
AL dans tous I	H' 30-44	1.144	0.066	1.062	0.028
AL dans tous I	H' 45-64	1.055	0.031	1.047	0.022
AL dans tous I	H' 65 +	1.068	0.056	1.054	0.027
AL dans tous I	F 0-14	1.148	0.062	1.064	0.027
AL dans tous I	F 15-29	1.126	0.054	1.112	0.028
AL dans tous I	F 30-44	1.134	0.057	1.064	0.027
AL dans tous I	F 45-64	1.068	0.041	1.049	0.025
AL dans tous I	F 65 +	0.948	0.021	0.992	0.018
AP dans tous I	H' 0-14	1.044	0.037	1.040	0.021
AP dans tous I	H' 15-29	1.148	0.064	1.103	0.025
AP dans tous I	H' 30-44	1.006	0.048	1.032	0.023
AP dans tous I	H' 45-64	1.036	0.017	1.034	0.014
AP dans tous I	H' 65 +	1.017	0.019	1.025	0.016
AP dans tous I	F 0-14	1.159	0.068	1.052	0.024
AP dans tous I	F 15-29	1.081	0.042	1.092	0.023
AP dans tous I	F 30-44	0.997	0.017	1.011	0.014
AP dans tous I	F 45-64	1.025	0.012	1.026	0.011
AP dans tous I	F 65 +	0.997	0.012	1.004	0.011

Nota: H: hispanique, L: locataire, I: ilot, H': Homme, F: femme, P: propriétaire, A: autre, NH: non hispanique, AS: asiatique.
(Exemple: LH dans IH: locataire d'origine hispanique vivant dans un ilot à majorité hispanique)

Le modèle de régression indique des taux de sous-dénombrement plus élevés pour les locataires que pour les propriétaires, peu importe l'origine ethnique. De plus, le groupe d'âge 15-29 présente un taux de sous-dénombrement beaucoup plus élevé que les autres groupes. Inversement, le groupe d'âge 45-64 est celui qui présente le taux le plus faible. La variable "sexe" n'étant pas statistiquement significative, elle n'a pas été incluse dans le modèle. Deux facteurs de redressement (hommes d'origine hispanique âgés de 65 et plus et vivant dans des ilots non hispaniques et locataires asiatiques de sexe masculin âgés de 65 et plus) avaient une variance estimée nulle et n'ont donc pas été inclus dans le modèle. Pour ces deux facteurs de redressement, le facteur fondé sur l'échantillon a tenu lieu de facteur prévu.

Tableau 6
Résultats du lissage des facteurs de redressement du TOR

Strates formées a posteriori	Sexe/Âge	Estimé		Prévu	
		Facteur de redressement (Y)	Erreur type	Facteur de redressement (AF)	Erreur type
LH dans IH	H' 0-14	1.131	0.020	1.130	0.016
LH dans IH	H' 15-29	1.247	0.030	1.211	0.021
LH dans IH	H' 30-44	1.165	0.029	1.144	0.020
LH dans IH	H' 45-64	1.099	0.043	1.114	0.024
LH dans IH	H' 65 +	1.055	0.044	1.110	0.023
LH dans IH	F 0-14	1.124	0.023	1.126	0.018
LH dans IH	F 15-29	1.234	0.032	1.203	0.022
LH dans IH	F 30-44	1.084	0.017	1.098	0.015
LH dans IH	F 45-64	1.125	0.040	1.121	0.024
LH dans IH	F 65 +	1.099	0.045	1.122	0.024
PH dans IH	H' 0-14	1.056	0.018	1.050	0.015
PH dans IH	H' 15-29	1.078	0.018	1.084	0.015
PH dans IH	H' 30-44	1.087	0.016	1.072	0.014
PH dans IH	H' 45-64	1.031	0.012	1.031	0.011
PH dans IH	H' 65 +	1.073	0.028	1.054	0.019
PH dans IH	F 0-14	1.059	0.020	1.051	0.016
PH dans IH	F 15-29	1.088	0.016	1.090	0.014
PH dans IH	F 30-44	1.033	0.012	1.034	0.011
PH dans IH	F 45-64	1.020	0.012	1.022	0.011
PH dans IH	F 65 +	1.033	0.019	1.035	0.015
H dans INH	H' 0-14	1.105	0.052	1.051	0.023
H dans INH	H' 15-29	1.154	0.054	1.106	0.025
H dans INH	H' 30-44	1.131	0.065	1.050	0.024
H dans INH	H' 45-64	1.063	0.050	1.036	0.023
H dans INH	H' 65 +	0.991	0.000	0.991	0.000
H dans INH	F 0-14	1.137	0.047	1.059	0.023
H dans INH	F 15-29	1.033	0.022	1.060	0.017
H dans INH	F 30-44	1.079	0.037	1.051	0.021
H dans INH	F 45-64	1.033	0.028	1.031	0.019
H dans INH	F 65 +	0.947	0.040	1.013	0.022
LAS dans tous I	H' 0-14	1.059	0.041	1.076	0.026
LAS dans tous I	H' 15-29	1.127	0.044	1.137	0.028
LAS dans tous I	H' 30-44	1.195	0.077	1.093	0.031
LAS dans tous I	H' 45-64	1.004	0.057	1.063	0.030
LAS dans tous I	H' 65 +	0.982	0.000	0.982	0.000
LAS dans tous I	F 0-14	1.067	0.047	1.079	0.028
LAS dans tous I	F 15-29	1.215	0.055	1.153	0.029
LAS dans tous I	F 30-44	1.173	0.105	1.087	0.032
LAS dans tous I	F 45-64	1.012	0.061	1.065	0.030
LAS dans tous I	F 65 +	1.212	0.127	1.087	0.032

asiatique, qui à leur tour ont un taux plus élevé que les autres. En ce qui concerne les enrégis-
traments erronés, le pourcentage de cas est plus élevé pour les locataires que pour les proprié-
taires alors que l'origine ethnique n'est pas une variable déterminante sur ce plan. Pour ce qui
a trait aux cas ayant fait l'objet d'une substitution, le pourcentage est plus élevé pour les loca-
taires d'origine hispanique et les autres locataires que pour les propriétaires d'origine hispa-
nique et les autres propriétaires. En revanche, on observe la situation inverse dans le cas des
Asiatiques.

Si nous examinons maintenant les taux de sous-dénombrement selon l'âge et le sexe, nous
constatons que le groupe d'âge 15-29 a le taux le plus élevé et que les hommes ont un taux plus
élevé que les femmes. Les groupes d'âge 0-14 et 30-44 présentent des taux de sous-dénombrement
comparables, lesquels sont légèrement au-dessous de la moyenne pour la région d'essai. Les
groupes d'âge 45-64 et 65 + présentent aussi des taux comparables, qui sont bien inférieurs
à ceux des autres groupes d'âge. Au point de vue de la distribution, ces résultats correspon-
dent assez bien à ceux observés antérieurement.

Le pourcentage de cas enregistrés incorrectement est le plus élevé pour les deux premiers
groupes d'âge (0-14 et 15-29). Les deux groupes suivants (30-44 et 45-64) présentent les pour-
centages les plus faibles. Curieusement, le groupe d'âge 65 + présente un pourcentage qui se
situe entre les pourcentages des groupes précités. Le pourcentage de cas enregistrés incorrec-
tement varie peu d'un sexe à l'autre. Enfin, pour ce qui a trait aux cas ayant fait l'objet d'une
substitution, les différences de pourcentage entre les groupes d'âge et entre les sexes sont faibles.

4.2 Estimations relatives aux petites régions

Comme il a été mentionné précédemment, avant d'opérer un redressement des données au
niveau de l'îlot, nous avons ajusté un modèle de régression dans le but de "lissier" les données
et de réduire les effets de la variabilité d'échantillonnage. Le modèle de régression a été ajusté
aux 70 facteurs de redressement qui avaient été définis par les strates formées a posteriori. Ce
genre de modèle sert à déterminer un mode de calcul du taux de sous-dénombrement uniforme
pour les régions ayant les mêmes caractéristiques. Dans un deuxième temps, on rajuste les fac-
teurs de redressement fondés sur l'échantillon de manière à les rendre comparables aux esti-
mations du modèle. Ce procédé s'inspire de l'estimateur de James-Stein et des estimateurs
empiriques de Bayes. Les variables indépendantes qui pouvaient être utilisées en l'occurrence
dans le modèle étaient des variables indicatrices pour les catégories "origine ethnique-mode
d'occupation", les groupes d'âge et les sexes. Il n'était pas permis d'inclure des termes d'inte-
raction dans le modèle. Le modèle qui pouvait être ajusté aux données et qui présentait des
coefficients significatifs (selon un modèle de régression non pondéré) était le suivant:

$$Y = 1.038 + .090(HR) + .044(AR) + .013(OR) + .058(A15-29) - .009(A45-64)$$

où Y = facteur de redressement axé sur le modèle

LH = 1 si locataire d'origine hispanique vivant dans un îlot à majorité hispanique
= 0 dans les autres cas

LAS = 1 si locataire d'origine asiatique vivant dans un îlot quelconque
= 0 dans les autres cas

AL = 1 si autre locataire vivant dans un îlot quelconque
= 0 dans les autres cas

A15-29 = 1 si groupe d'âge 15-29
= 0 dans le cas contraire

A45-64 = 1 si groupe d'âge 45-64
= 0 dans le cas contraire.

Taux de sous-dénombrement (en pourcentage) et composantes de l'estimateur bivalent pour les strates formées a posteriori

Strates formées a posteriori	Taux de sous-dénombrement ^a	Pourcentage de cas de l'échantillon P non apparés	Pourcentage de cas de l'échantillon D enregistrés incorrectement ^b	Pourcentage de cas ayant fait l'objet d'une substitution dans le recensement
------------------------------	----------------------------------------	---------------------------------------------------	-------------------------------------------------------------------------------	------------------------------------------------------------------------------

Locataires d'origine hispanique vivant dans des flots à majorité hispanique	13.7	17.1	2.6	1.7
Propriétaires d'origine hispanique vivant dans des flots à majorité hispanique	5.5	8.1	1.2	1.5
Personnes d'origine hispanique vivant dans des flots à minorité hispanique	7.5	9.7	1.4	1.3
Locataires d'origine asiatique	11.1	13.4	2.1	1.2
Propriétaires d'origine asiatique	4.6	6.8	1.2	1.5
Autres locataires	9.9	12.9	2.4	1.7
Autres propriétaires	3.8	5.8	1.3	0.9
0-14	8.8	11.9	2.2	1.6
15-29	13.6	16.2	2.1	1.6
30-44	8.6	10.8	1.4	1.4
45-64	4.5	6.6	1.3	1.4
65 +	3.3	5.9	1.7	1.3
Homme	9.7	12.1	1.7	1.5
Femme	8.3	10.8	1.9	1.5
Total	9.0	11.4	1.8	1.5

^a La sommation de toutes les estimations est étendue à toutes les autres catégories.

^b Les cas enregistrés incorrectement comprennent les cas non apparables qui n'ont pas fait l'objet d'une substitution.

enregistrés incorrectement (100(EE/REN)) et le pourcentage de cas ayant fait l'objet d'une substitution (100(SUB/REN)).

Une caractéristique de l'estimateur bivalent veut que la somme des estimations pour plusieurs catégories n'égal pas l'estimation directe de la somme pour l'ensemble de ces catégories. Ainsi, pour assurer la cohérence des estimations indiquées dans le tableau 5, on a étendu la sommation de toutes les estimations aux autres catégories pertinentes.

Si nous examinons les taux de sous-dénombrement indiqués dans le tableau 5 pour les catégories "origine ethnique-mode d'occupation", nous en concluons que le mode d'occupation est une bonne variable de stratification, les locataires ayant un taux de sous-dénombrement plus élevé que les propriétaires; l'origine ethnique influe aussi sur le taux de sous-dénombrement, les personnes d'origine hispanique ayant un taux plus élevé que les personnes d'origine

À la fin de la période d'interview, 5,714 (93,2%) unités de logement avaient fait l'objet d'une interview complète avec un membre du ménage, 193 (3,1%) avaient été classées parmi les logements inoccupés et 189 (3,1%) avaient fait l'objet d'une interview complète avec une personne qui n'était pas membre du ménage (par exemple, un voisin). Seulement 32 (0,5%) unités de logement ont été classées parmi les cas de non-interview. Ce taux extrêmement faible s'explique par le fait que l'on avait alloué cinq semaines pour les interviews.

Une fois remplis, les questionnaires de l'EP ont été préparés en vue de l'appariement informatisé avec le fichier du recensement. L'appariement informatisé a été divisé en deux phases: la première visait à appairer les données de l'EP avec celles de l'échantillon D et la seconde visait à appairer tous les cas de l'échantillon F qui n'avaient pu l'être dans la première phase avec les autres données du recensement. La seconde phase de l'appariement informatisé servait aussi à appairer les personnes ayant déménagé entre le jour du recensement et le jour de l'interview pour l'EP avec les erreurs de géocodage (c.-à-d., lorsque l'unité de logement est associée au mauvais lot). La première phase de l'appariement informatisé a permis de déterminer une concordance pour 14,700 cas (73,5%) de l'échantillon F et une concordance probable pour 2,550 autres cas (12,0%). La seconde phase de l'appariement informatisé a permis de déterminer une concordance pour 130 personnes (0,7%) et une concordance probable pour 570 autres personnes (2,9%). Étant donné le faible pourcentage de cas de l'échantillon F pour lesquels la seconde phase de l'appariement a permis de déterminer une concordance, nous en avons conclu que le géocodage dans Los Angeles comportait peu d'erreurs.

L'appariement manuel a servi à revoir les résultats de l'appariement informatisé. Il a servi aussi à identifier les cas qui n'avaient pu être appariés à cause d'un manque de données (pour ces cas, il faut recourir à l'imputation). Enfin, c'est durant l'appariement manuel qu'ont été rédigés les questionnaires de suivi pour les cas non résolus des échantillons F et D. Le suivi sur place visait 1,551 unités de logement, dont 1,511 (97,4%) avaient fait l'objet d'une interview complète. Cette opération a été suivie d'un appariement final. Les résultats de l'échantillon F indiquent que 17,018 (85,2%) personnes ont pu être appariées à un enregistrement du recensement tandis que 2,373 (11,9%) n'ont pu l'être. En outre, 426 (2,1%) personnes ont été considérées comme exclues du champ de l'enquête (il s'agissait pour la plupart de personnes qui demeuraient à l'extérieur de la région d'essai le jour du recensement) et 161 (0,8%) personnes représentaient des cas non résolus (auxquels on a imputé ultérieurement un code d'appariement). Les résultats définitifs de l'échantillon D indiquent que 19,637 (93,6%) personnes ont été enregistrées correctement dans le recensement tandis que 360 (1,7%) l'ont été incorrectement. De plus, 976 (4,7%) personnes représentaient des cas non résolus auxquels on a imputé un code de dénombrement.

Après l'appariement final, toutes les données manquantes, y compris les codes d'appariement pour l'échantillon F et les codes de dénombrement pour l'échantillon D, ont été imputées. Les résultats ont servi à calculer les estimations bivalentes. Les estimations ont été lissées puis reproduites au niveau de l'lot pour créer un fichier du recensement redressé. La réduction du délai de production des estimations du taux de sous-dénombrement a été surtout attribuable aux opérations d'appariement. Pour le TOR, l'appariement informatisé et l'appariement manuel se sont étendus sur environ trois mois tandis que les opérations d'appariement pour l'EP de 1980 s'étaient déroulées sur plus d'un an. Une meilleure planification des opérations ainsi qu'une facilité accrue pour consulter les documents du recensement sont deux facteurs qui ont aussi contribué à réduire les délais de production.

4. ESTIMATIONS

4.1 Estimations relatives aux strates formées a posteriori

Cette sous-section présente les estimations du taux de sous-dénombrement pour divers groupes de strates formées a posteriori. Le tableau 5 donne le taux de sous-dénombrement $(100(1-R_{EN}/EB))$, le pourcentage de cas non appariés $(100(1-M/N_p))$, le pourcentage de cas

Tableau 4
Calendrier des opérations pour le TOR de 1986

Opération	Début	Fin
Jour du recensement	16 mars	16 mars
Suivi des cas de non-réponse	9 avril	8 mai
Saisie des noms du recensement	23 mai	10 juin
Fichier du recensement pour l'appariement	8 août	15 août
Liste d'adresses pour l'EP	17 juin	21 juin
Sous-échantillonnage pour l'EP	25 juin	1 juillet
Période d'interview pour l'EP	25 juin	8 août
Saisie des données de l'EP	21 juillet	19 août
Appariement informatisé	28 août	9 septembre
Appariement informatisé (deuxième phase)	9 septembre	3 octobre
Appariement manuel	15 septembre	31 octobre
Suivi sur place	23 septembre	6 novembre
Appariement des résultats du suivi	29 septembre	6 novembre
Saisie des résultats de l'appariement	21 octobre	10 novembre
Préparation des fichiers des échantillons P et D	11 novembre	2 janvier
Imputation	5 janvier	11 janvier
Version finale du fichier du recensement	—	5 janvier
Estimation relative aux strates formées à posteriori	12 janvier	11 février
Estimation relative aux petites régions	12 février	22 février

Le recensement a été réalisé à l'aide d'un questionnaire qui a été posté à chaque unité de logement connue et qu'un membre du ménage devait remplir le jour du recensement (16 mars). Les ménages qui négligeaient de retourner leur questionnaire recevaient la visite d'un agent recenseur qui procédait à une interview sur place. C'est ce qu'on appelle le suivi des cas de non-réponse. Une fois remplis, les questionnaires étaient acheminés au service de traitement pour la saisie des données, qui comprenaient en l'occurrence tous les noms du recensement. La première étape de l'EP consistait à dresser une liste indépendante de toutes les adresses figurant dans les lots échantillonnés. On a confronté la liste en question avec une liste administrative pour s'assurer de son exactitude et de son intégralité. Cette mesure de contrôle de la qualité a permis de constater que 127 lots (67%) ne requéraient aucune modification à la liste d'adresses. Les 63 autres lots (33%) ont dû faire l'objet de modifications par l'intermédiaire du contrôle de qualité puis les adresses correspondantes ont été relistées. Cette opération a permis d'ajouter des adresses dans 37 lots, d'en corriger dans 39 lots et d'en soustraire dans 9 lots. (Comme certains lots ont fait l'objet de plus d'un genre de modifications, la somme des nombres ci-dessus ne correspond pas au nombre total d'lots qui ont dû faire l'objet de modifications.) La mesure de contrôle de la qualité n'a entraîné que des corrections mineures pour les autres adresses. À l'issue de cette opération de contrôle, tous les lots comptant 70 unités de logement ou plus ont été divisés par sous-échantillonnage à l'aide des côtes d'lots ou des tranches d'adresses.

L'interview de l'EP a été réalisée sur place. Elle visait à recueillir des renseignements sur les caractéristiques démographiques de toutes les personnes qui demeuraient à l'endroit visité le jour de l'enquête. L'interview comportait des questions particulières comme celles portant sur le domicile le jour du recensement, l'adresse postale, d'autres lieux d'habitation comme les foyers d'étudiants, et les autres personnes qui pouvaient avoir demeuré à l'endroit visité le jour du recensement. Une mesure de contrôle de la qualité du questionnaire de l'EP a permis de vérifier la liste des noms. Quatre-vingt-seize pour cent des questionnaires vérifiés ont passé le test. Pour ce qui est des autres (4%), on a dû procéder à une nouvelle interview et corriger les questionnaires en conséquence.

redressement fondé sur l'échantillon et du facteur de redressement prévu obtenir le facteur de redressement

(5)
$$AF_i = \left(Y_i/\sigma_i^2 + \sum_j^j X_{ij}B_j/\epsilon^2 \right) \left(\sigma_i^{-2} + \epsilon^{-2} \right)^{-1},$$

qui sert à corriger les données de recensement relatives à l'îlot. La variance de AF_i peut être déduite des résultats publiés dans Freedman et Navidi (1986).
L'estimation synthétique a servi à reproduire le redressement effectué dans chaque strate au niveau de l'îlot. L'estimateur synthétique est défini

(6)
$$ADJ_{ij} = AF_i \times CEBN_{ij},$$

où i et j désignent respectivement la strate formée a posteriori et l'îlot et ADJ est le chiffre de population redressé pour l'îlot.

Le chiffre de population redressé ADJ_{ij} est en règle générale un nombre fractionnaire. Or, le recensement sert à dénombrer des unités. Par conséquent, si nous voulons établir une comparabilité entre les opérations de redressement et le recensement, il est nécessaire de convertir les nombres fractionnaires en nombres entiers. Cette conversion (appelée arrondissement continu) permet d'arrondir vers le haut ou vers le bas tous les nombres fractionnaires (Caussey et coll. 1985).

Après avoir arrondi les estimations redressées des îlots, on a produit des chiffres selon l'âge, l'origine ethnique et le sexe qui indiquaient le nombre de personnes qu'il fallait ajouter ou retrancher dans chaque îlot. Dans le cas d'un sous-dénombrement, on prélevait aléatoirement dans l'îlot un enregistré qui présentait la même série de caractéristiques que la personne oubliée et on en tirait un nouvel enregistré identique. On a utilisé une catégorie n'ayant pas trait aux ménages pour ajouter des personnes dans les listes du recensement de manière qu'il ne soit pas nécessaire de tenir compte des rapports entre les membres du ménage ou de créer de nouveaux ménages. Zaslavasky (1988) présente une autre méthode, fondée sur la pondération, pour ajouter des personnes et des ménages dans les îlots de recensement. Dans le cas d'un surdénombrement, on signalerait les enregistrements qui présentent les caractéristiques voulues et on les incluerait dans les totalisations corrigées du recensement.

3. OPÉRATIONS ET CALENDRIER DES OPÉRATIONS

Le recensement d'essai avait principalement pour but d'analyser le calendrier et les caractéristiques des opérations de redressement des données du recensement. Les EP qu'avait réalisées le Censur Bureau dans les années antérieures s'étaient étendues sur au moins deux ans. Par exemple, l'EP de 1980 a produit des estimations du taux de sous-dénombrement à l'autisme de 1981 puis une série d'estimations finales au début de 1982.
Le tableau 4 énumère les principales opérations du recensement et de l'EP ainsi que les dates qui marquent le début et la fin de chacune de ces opérations. On remarquera qu'il s'écoule un certain temps entre la fin d'une opération et le début d'une autre car les opérations du recensement et de l'EP ne sont pas toutes indiquées. Certaines périodes se chevauchent car les opérations correspondantes se déroulaient simultanément. On a amorcé l'EP une fois que les principales opérations régionales touchant le recensement ont été terminées, cela ayant pour but d'éviter que le personnel des bureaux régionaux n'ait à exécuter simultanément les opérations des deux enquêtes.

part. Le tableau 3 donne les sept catégories "origine ethnique — mode d'occupation" qui ont été subdivisées selon l'âge (0-14, 15-29, 30-44, 45-64, 65 +) et le sexe pour former les 70 strates qui ont servi à l'estimation.

Le tableau 3 indique également la taille des échantillons P et D. La taille de l'échantillon P est moins élevée que celle de l'échantillon D à cause notamment des membres de l'échantillon P qui sont venus s'établir dans la région d'essai après le recensement et qui sont exclus par conséquent du champ de l'enquête.

2.5 Traitement des données manquantes

Le calcul des estimations bivariantes exige un fichier de données complet. Or, comme dans le cas de n'importe quelle enquête par sondage, le test de 1986 n'a pas permis de recueillir toutes les données voulues. Schenker (1988) décrit les méthodes utilisées pour résoudre le problème des données manquantes et analyse notamment les effets de diverses hypothèses concernant les données manquantes sur les estimations bivariantes. Afin d'avoir une idée d'ensemble de la question, nous décrivons brièvement ici ces méthodes.

Dans le test de 1986, il manquait des données sur les caractéristiques des personnes et du ménage, le code d'appariement (apparié/non-apparié) pour les personnes de l'échantillon P et le code de dénombrement (exact/erroné) pour les personnes de l'échantillon D. Pour les cas de non-interview de l'échantillon P, on a eu recours à un redressement par pondération. L'imputation des caractéristiques manquantes s'est faite à l'aide d'une méthode "hot-deck". En ce qui concerne les codes d'appariement, on s'est servi d'un modèle de régression logistique pour estimer la probabilité d'appariement. Au lieu d'attribuer un code d'appariement définitif à chaque cas non résolu, on s'est servi des probabilités estimées dans le calcul des estimations bivariantes. On a procédé de façon analogue pour les codes de dénombrement manquants de l'échantillon D.

2.6 Estimation relative aux petites régions

Afin que les utilisateurs disposent de données redressées à tous les niveaux d'aggrégation, on calcule des estimations du taux de sous-dénombrement jusqu'au niveau de l'ilot, qui est la plus petite unité géographique. Avant d'effectuer ces calculs toutefois, on utilise un modèle de régression pour "lisser" les effets de l'erreur d'échantillonnage. Dans ce modèle, des facteurs de redressement servent de variable dépendante. Un facteur de redressement est défini comme le quotient de l'estimateur bivalent par le chiffre du recensement:

(3)
$$Y = EB/REN,$$

où REN et EB sont définis comme auparavant.

Le modèle de régression s'exprime par le formule suivante:

(4)
$$Y_i = B_0 + B_1X_{i1} + \dots + B_pX_{ip} + S_i + E_i,$$

où Y_i est le facteur de redressement pour la i -ième strate formée a posteriori ($i = 1, \dots, 70$), X_{ij} est la variable indépendante ($j = 1, \dots, p$), B_j est le coefficient de régression à estimer, S_i est l'erreur d'échantillonnage du facteur de redressement, E_i est l'erreur du modèle et les valeurs de S_i et de E_i sont indépendantes et distribuées selon une loi normale avec une moyenne nulle et des variances égales à σ_i^2 et à ϵ^2 respectivement. Les valeurs de B_j sont estimées à l'aide des méthodes du maximum de vraisemblance (Erickson et Kadane 1985). Les valeurs de σ_i^2 sont estimées directement à l'aide de l'échantillon. On fait la moyenne du facteur de

6000 unités de logement. Le tableau 2 donne la répartition des îlots échantillonnés selon les strates. Les grands îlots qui comptent 70 unités de logement ou plus ont fait l'objet d'un sous-échantillonnage afin de réduire la tâche des intervieweurs. Le sous-échantillonnage consistait à diviser l'îlot en grappes de 35 à 50 unités de logement à l'aide de tranches d'adresses ou de côtés d'îlot. On choisissait une grappe au hasard dans le but d'interviewer des membres de l'échantillon P. Quant à l'échantillon D, il comprenait toutes les personnes qui faisaient partie de cette grappe le jour du recensement.

2.4 Stratification a posteriori

L'estimateur bivalent est biaisé et ce biais peut être élevé s'il y a de fortes différences de taux de sous-dénombrement entre les sous-groupes de la population (Wolter 1986). Pour réduire au minimum ce biais, on a stratifié a posteriori la région d'essai selon des critères précis de manière à ce que les taux de sous-dénombrement soient comparables dans toutes les strates. On a ensuite calculé des estimations bivalentes dans chaque strate ainsi formée.

Les strates ont été choisies par suite d'un examen de la structure de la région d'essai et d'une analyse des données de l'EP de 1980. On a observé que l'origine ethnique était la variable qui influait le plus sur le taux de sous-dénombrement. On a donc défini trois groupes ethniques: les personnes d'origine hispanique, les personnes d'origine asiatique et les autres. Il n'a pas été possible de créer une strate pour les personnes de race noire puisque la région d'essai comptait très peu de noirs. Le fait d'être locataire constituait une variable explicative importante dans notre étude précédente (Isaki et coll. 1987). Par conséquent, le mode d'occupation a aussi été considéré dans la construction des strates. Par ailleurs, comme on croyait que les personnes d'origine hispanique vivant dans des îlots qui comptaient moins de 50% de personnes d'origine hispanique (appelés îlots à minorité hispanique) n'avaient pas le même taux de sous-dénombrement que les autres personnes de même origine, on les a incluses dans une strate à

Tableau 3

Catégories "origine ethnique — mode d'occupation"
Echantillon P Echantillon D
utilisées dans la stratification a posteriori, y compris la taille des échantillons

Localitaires d'origine hispanique	8,182	8,739
vivant dans des îlots à majorité hispanique		
Propriétaires d'origine hispanique		
vivant dans des îlots à majorité hispanique		
hispanique	5,688	5,867
Personnes d'origine hispanique		
vivant dans des îlots à minorité hispanique		
hispanique	896	1,005
Localitaires d'origine asiatique	666	911
Propriétaires d'origine asiatique	1,144	1,230
Autres localitaires	1,135	1,316
Autres propriétaires	1,841	1,908
Total	19,552	20,976

Tableau 2

Strates d'échantillonnage et répartition des ilots échantillonnés	
Strates d'échantillonnage	Nombre d'ilots échantillonnés
1. Ilots à majorité hispanique constitués de grands immeubles à logements multiples	8
2. Ilots à majorité hispanique constitués de petits immeubles à logements multiples	49
3. Ilots à majorité hispanique constitués de maisons familiales	39
4. Ilots à majorité asiatique	35
5. Autres ilots	38
6. Ilots comptant deux unités de logement ou moins	21

peuvent être classés avec certitude parmi les concordances ou les non-concordances, nous avons aussi soustrait des chiffres du recensement le nombre de personnes non appartiables. Toutes les personnes correspondantes de l'échantillon P ont été désignées comme des non-concordances et classées dans la case N₂₁.

2.3 Plan de sondage

Le plan de sondage prévoyait un échantillonnage stratifié où l'ilot servait d'unité d'échantillonnage. Deux catégories de données ont servi à stratifier la région d'essai: le nombre d'unités de logement par ilot, tiré du fichier d'adresses du recensement de 1986, et les résultats de l'application des données du recensement de 1980 sur l'origine ethnique aux unités géographiques du recensement de 1986. Comme cette opération ne pouvait s'effectuer qu'au niveau du secteur de recensement, qui peut contenir de un à six ilots, la répartition des groupes ethniques s'est faite à ce niveau. Tous les ilots d'un secteur de recensement ont été identifiés au même groupe ethnique et se sont trouvés par conséquent dans la même strate.

La région d'essai a été divisée en six strates d'échantillonnage comme l'indique le tableau 2. Tous les ilots qui comprenaient des lieux particuliers (le plus souvent des foyers collectifs) ont été classés dans une strate d'échantillonnage distincte. Ces ilots étaient considérés comme exclus du champ de l'enquête et n'ont pas été échantillonnés. Les petits ilots ont été classés dans une strate distincte afin de réduire la variance d'échantillonnage. Tous les ilots des secteurs de recensement qui comprenaient au moins 18% d'Asiatiques ont servi à former la strate à majorité asiatique. Tous les ilots à majorité non asiatique des secteurs de recensement qui comprenaient au moins 40% de personnes d'origine hispanique ont servi à former les trois strates à majorité hispanique. Les ilots qui n'étaient pas inclus dans l'une ou l'autre des strates précédentes ont servi à former la strate réservée aux personnes d'autres origines.

Les données de 1986 sur le logement renfermaient aussi de l'information sur les maisons familiales et les immeubles à logements multiples. Ces données ont servi à partager les strates à majorité hispanique en trois catégories: maisons familiales, petits immeubles à logements multiples et grands immeubles à logements multiples. La strate de la troisième catégorie était composée des ilots à majorité hispanique où au moins la moitié des unités de logement se trouvaient dans des immeubles de dix logements ou plus. La strate de la première catégorie était formée des ilots à majorité hispanique qui comprenaient plus de 50% de maisons familiales. Enfin, la strate de la seconde catégorie comprenait tous les autres ilots à majorité hispanique. Des ilots ont été prélevés dans chacune des strates au moyen d'un échantillonnage systématique avec probabilités égales. L'échantillon était composé de 190 ilots renfermant environ

gécodage, la répétition, l'information fausse, les personnes nées après le jour du recensement, les personnes décédées avant le jour du recensement et les cas non identifiables. On dit qu'il y a erreur de gécodage lorsqu'un enrégistrement du recensement existe à l'extérieur du secteur de recherche ou de la région d'essai. Les cas non identifiables sont des enrégistrement sans nom. Ils entraînent une surestimation du nombre d'enrégistrement erronés mais sont traités comme ceux-ci dans le calcul de l'estimateur.

2.2 Estimation bivalente

Afin d'estimer la population totale, on utilise un estimateur bivalent qui réunit les données des échantillons P et D. Wolter (1986) décrit divers estimateurs bivalents et les hypothèses sur lesquelles ils reposent. L'estimateur bivalent utilisé dans le TOR est défini

$$(1) \quad EB = \frac{N_p(REN-SUB-EE)}{M},$$

où N_p est l'estimateur de la population totale de l'EP, REN est le chiffre du recensement non redressé, SUB est le nombre de substitutions de personnes dans le recensement, EE est l'estimateur du nombre d'enrégistrement erronés et de personnes non identifiables dans le recensement selon l'échantillon D et M est l'estimateur du nombre de personnes qui font partie à la fois de la population du recensement et de celle de l'EP. On parle de substitution de personne lorsqu'une personne quelconque a été enrégistrée dans le recensement avec moins de deux caractéristiques démographiques. Afin de mieux illustrer quelques-unes des caractéristiques uniques de l'estimateur bivalent, nous exposons dans le tableau ci-dessous la classification des membres de la population.

Les effectifs du tableau 1 sont estimés à l'aide des composantes de l'estimateur bivalent: $N_{11} = M, N_{+1} = N_p, N_{1+} = REN-SUB-EE$. La valeur de N_{22} est inobservable par définition mais on peut l'estimer en supposant que le recensement et l'échantillon P de l'EP sont indépendants l'un de l'autre. L'estimation de N_{22} est définie par l'équation suivante:

$$(2) \quad N_{22} = N_{12}N_{21}/N_{11}.$$

A l'aide des estimateurs définis ci-dessus, nous estimons la population totale par $N_{++} = EB$. A cause des problèmes que soulève l'appariement des données de recensement, il faut effectuer des opérations particulières pour ne pas surestimer la population. L'estimateur bivalent suppose que chaque personne est classée dans une seule case du tableau 1. Ainsi, au lieu d'utiliser simplement les chiffres du recensement, nous soustrayons de ces chiffres l'estimation du nombre d'enrégistrement erronés pour obtenir une estimation du nombre **réel** de personnes dénombrées dans le recensement. En outre, l'estimateur bivalent suppose que chaque personne peut être désignée comme une concordance ou une non-concordance. Comme les enrégistrement du recensement qui ne renferment pas assez d'information pour les besoins de l'appariement (par exemple, absence de nom ou moins de deux caractéristiques démographiques) ne

Tableau 1

Classification bivalente

Population cible de l'échantillon P			
Inclus		Exclus	
Recensement	Inclus	Exclus	Total
Ennumération			
	N_{11}	N_{12}	N_{1+}
	N_{21}	N_{22}	N_{2+}
	N_{+1}	N_{+2}	N_{++}
Total			

Dans la Section 2, nous exposons la méthode utilisée en 1986 pour estimer les taux de sous-dénombrement et indiquons comment ces estimations servent au redressement des chiffres du recensement. Dans la Section 3, nous analysons le calendrier des opérations prévu pour le TOR, y compris les opérations sur le terrain et l'appariement. La Section 4 contient un résumé des estimations des taux de sous-dénombrement pour les strates formées a posteriori et les ilots. Enfin dans la Section 5, nous faisons un résumé des principaux résultats de notre analyse et tirons quelques conclusions.

2. MÉTHODOLOGIE

2.1 Aperçu des échantillons utilisés dans l'estimation

Pour estimer la population, on a utilisé deux échantillons dans l'EP: l'échantillon P (pour population) et l'échantillon D (pour dénombrement). Le premier sert à mesurer les omissions du recensement tandis que le second sert à mesurer les enregistrements erronés du recensement. L'échantillon P est un échantillon d'ilots qui s'accompagne d'un listing indépendant d'unités de logement et d'interviews sur place tandis que les données de l'échantillon D sont les enregistrements (chiffres) du recensement qui se rapportent aux ilots échantillonnés. L'échantillon P a servi à recueillir les données nécessaires à l'appariement et à l'estimation, y compris le domicile au jour du recensement. Pour les besoins du plan de sondage, il a fallu définir les éléments de l'échantillon P. L'échantillon P était constitué de toutes les personnes qui demeuraient aux adresses échantillonnées au moment de l'EP. Nous aurions pu choisir d'interviewer les personnes qui demeuraient à ces adresses le jour du recensement. Nous avons éliminé cette possibilité car tous les cas de déménagement impliquent une interview par personne interposée (c'est-à-dire une interview réalisée auprès d'une personne qui n'est pas membre du ménage). Selon la méthode que nous avons adoptée, des personnes ayant déménagé pouvaient se trouver à l'adresse échantillonnée et on pouvait les interviewer sans avoir affaire à un enquête-substitut. Cependant, toutes les personnes qui ont quitté la région d'essai entre le jour du recensement et le jour de l'EP n'avaient aucune chance de faire partie de l'échantillon P. Par ailleurs, toutes les personnes de l'échantillon P qui demeuraient à l'extérieur de la région d'essai le jour du recensement étaient considérées comme exclues du champ de l'enquête.

Après l'interview, toutes les personnes de l'échantillon P ont été appariées à des enregistrements du recensement, l'appariement se faisant par ordinateur avec une vérification manuelle. Il a aussi fallu définir pour les besoins du plan de sondage l'étendue du secteur de recherche pour l'appariement. Selon l'EP, une personne de l'échantillon P était considérée comme appariée si elle avait été enregistrée lors du recensement n'importe où dans la région d'essai. On aurait pu définir un secteur de recherche moins étendu comme l'ilot servant à l'EP et les ilots avoisinants. Ainsi, il y a concordance pour une personne de l'échantillon P seulement lorsque si la personne correspondante dans le recensement se trouve dans ce secteur de recherche. À ce propos, l'EP de 1990 utilisera un secteur de recherche restreint pour l'appariement.

Tous les cas qui n'ont pu être appariés ont dû faire l'objet d'un suivi qui visait à recueillir des renseignements additionnels pour l'appariement. La tâche des personnes affectées au suivi de l'échantillon P a été sensiblement allégée du fait que toutes les questions nécessaires à l'appariement avaient été posées au moment de l'interview initiale. Par conséquent, seuls les cas où il manquait de l'information sur les caractéristiques personnelles, les cas où l'adresse de personnes ayant déménagé était incomplète et les cas d'appariement incertains ont dû faire l'objet d'un suivi. Les cas non appariés de l'échantillon P ont été jugés résolus et n'ont plus été renvoyés à l'équipe chargée du suivi. De nombreuses personnes de l'échantillon D ont pu être appariées à des personnes de l'échantillon P et ces cas ont été résolus sans devoir réaliser une autre interview. Tous les cas de l'échantillon D qui n'ont pu être résolus à l'aide de l'interview de l'échantillon P ont dû faire l'objet d'une interview de rappel qui visait à déterminer leur code de dénombrement. Les opérations sont analysées plus en détail à la section suivante. Les types d'enregistrements erronés que sert à mesurer l'échantillon D comprennent l'erreur de

dénumbrément sont plus élevés dans les grandes villes que dans les régions rurales et plus élevés chez les hommes que chez les femmes. Les membres du groupe d'âge 20-45 présentent aussi un fort taux de sous-dénumbrément.

Depuis 1950, on mesure le taux de sous-dénumbrément aux États-Unis au moyen d'une enquête post-censitaire (EP) et d'une analyse démographique. Le Census Bureau a laissé savoir que ces deux méthodes seraient les principaux moyens d'estimer le taux de sous-dénumbrément pour le recensement de 1990. Une EP utilise un échantillon indépendant de personnes qui sont apparues aux enregistrements du recensement afin d'estimer la population totale. Marks (1978) et le U.S. Census Bureau (1979) décrivent les expériences qui ont déjà été faites au moyen d'une EP pour mesurer le taux de couverture du recensement. Par ailleurs, l'analyse démographique fait appel aux registres des naissances et des décès et à d'autres fichiers administratifs pour estimer la population totale des États-Unis. Fay et coll. (1988) présentent les estimations des taux de sous-dénumbrément de 1980 qui ont été établies à l'aide de l'analyse démographique et du programme post-censitaire (PP).

En 1980, une analyse plus approfondie des chiffres du recensement a donné lieu à un certain nombre de poursuites judiciaires par lesquelles on exigeait un redressement de ces chiffres. Ces poursuites avaient été engagées pour des raisons comme la différence de taux de sous-dénumbrément entre les noirs et les non-noirs, la création dans les années 1970 d'un mécanisme de partage des revenus qui liait directement des fonds publics aux chiffres de population et la baisse de population dans des villes et des États pour lesquels le taux de sous-dénumbrément avait toujours été supérieur à la moyenne. Le U.S. Census Bureau s'est refusé au redressement des chiffres de 1980 en faisant valoir que le calcul du taux de sous-dénumbrément était exposé aux erreurs et qu'un redressement des chiffres du recensement ne serait pas vraiment utile.

Le Census Bureau s'est néanmoins engagé après 1980 dans un programme de recherche visant à étudier diverses façons d'améliorer le calcul des taux de sous-dénumbrément (Mulry et coll. 1981 et Hogan 1984). À ce sujet, Hogan (1984) a proposé une série de tests à effectuer conjointement avec les recensements d'essai. Cela a commencé par une EP à Tampa (Floride) en 1985 dans le but de tester et d'évaluer l'appariement par ordinateur. Ce test a confirmé la faisabilité de ce mode d'appariement (Jaro et Childers 1986). Des recensements d'essai et des EP ont aussi été réalisés en 1986 à Los Angeles et au Mississippi. Dans le premier cas, l'EP visait à tester le calendrier et les caractéristiques des opérations de redressement des chiffres du recensement. Dans le second cas, l'EP visait à évaluer les opérations d'une telle enquête dans une région d'essai rurale (Anolik 1988).

Une enquête pré-censitaire (EPR) a aussi été réalisée à Los Angeles en 1986 pour déterminer s'il était possible de réduire le délai des opérations de redressement en exécutant une partie du travail sur le terrain avant le recensement plutôt qu'après comme dans une EP (Wolfgang 1987). En 1987, on a réalisé une enquête pré-censitaire dans le Dakota du Nord afin d'évaluer les opérations d'une telle enquête dans des régions rurales; dans ce cas toutefois, on a procédé à des interviews sur place au lieu d'envoyer des questionnaires par la poste comme cela s'était fait dans les autres régions d'essai. Enfin, on est à préparer la répétition générale de 1988. Cette répétition générale permettra de tester toutes les opérations de recensement avant l'exécution du recensement décennal de 1990.

Cet article porte plus spécialement sur l'EP de 1986 pour le Central Los Angeles County, appelée Test des opérations de redressement (TOR) et réalisée conjointement avec le recensement d'essai. La région d'essai comprenait trois grands groupes ethniques: les personnes d'origine hispanique (qui constituaient environ 75% de la population totale), les personnes d'origine asiatique (environ 15% de la population totale) et les autres personnes, en majorité de race blanche (environ 10% de la population totale). L'EP a révélé un taux de sous-dénumbrément estimé de 9%. Pour les trois principaux groupes ethniques de la région d'essai, les taux de sous-dénumbrément étaient de 9,8% pour les personnes d'origine hispanique, de 7,3% pour les personnes d'origine asiatique et de 6,2% pour les autres. Dans cet article, nous décrivons la manière dont nous estimons ces taux.

Test des opérations de redressement de 1986 dans le Central Los Angeles County

GREGG DIFFENDAL¹

RÉSUMÉ

En vue du recensement décennal de 1990, le Census Bureau a étudié la possibilité de redresser les chiffres du recensement pour tenir compte du taux de sous-dénombrement estimé. À cette fin, il a exécuté un recensement d'essai dans un secteur à majorité hispanique du Central Los Angeles County afin de vérifier le calendrier et les caractéristiques des opérations de redressement réalisées au moyen d'une enquête post-censitaire (EP). Cet article vise à exposer les méthodes qui ont été utilisées pour produire des données de recensement qui tiennent compte de la population non dénombrée; on y trouve également les résultats de l'application de ces méthodes. Les méthodes utilisées pour le redressement des données de recensement comprenaient l'élaboration d'un plan de sondage, l'estimation "bivalente" et le calcul d'estimations régionales. Le plan de sondage prévoyait un échantillon d'îlots stratifiés selon l'origine ethnique. L'appariement s'est fait par ordinateur tandis que les opérations de contrôle et de résolution ont été exécutées manuellement. L'estimateur "bivalent", appelé aussi estimateur de Petersen ou saisie-resaisie, a servi à estimer la population. À cause de la nature des recensements, les chiffres du recensement ont été redressés avant de servir au calcul de l'estimateur bivalent. Avant de corriger les estimations régionales, on a ajusté un modèle de régression au facteur de redressement (estimateur divisé par le chiffre du recensement) afin de réduire les effets de la variabilité d'échantillonnage. Un estimateur synthétique a permis d'effectuer le redressement jusqu'au niveau de l'îlot. Les résultats de l'estimation bivalente sont présentés pour la région d'essai selon les trois principaux groupes ethniques (hispanique, asiatique, autre), le mode d'occupation, l'âge et le sexe. Enfin, nous présentons en bref les résultats du redressement des estimations régionales du recensement par îlot et nous en faisons l'analyse.

MOTS CLÉS: Sous-dénombrement dans le recensement; estimation bivalente; estimation synthétique; enquête post-censitaire.

1. INTRODUCTION

Depuis le premier recensement de la population des E.-U. en 1790, on a toujours eu de la difficulté à dénombrer toutes les personnes qui devaient l'être. Grâce aux progrès réalisés en démographie et en statistique, on a pu établir à partir de 1950 des estimations du taux de couverture du recensement. Ces estimations ont servi à évaluer les faiblesses du recensement et à définir les points qui devaient être améliorés en vue des recensements subséquents. Les estimations du taux de couverture du recensement ont connu une amélioration soutenue depuis la publication des premières estimations du genre en 1950. Une série d'estimations indique que le taux de sous-dénombrement pour l'ensemble du pays était de 4,4% en 1950, de 3,3% en 1960, de 2,8% en 1970 et de 1% en 1980. Malgré cette baisse soutenue du taux de sous-dénombrement, les estimations demeurent élevées pour certains groupes. Dans le cas des personnes de race noire, par exemple, le taux de sous-dénombrement se maintient à environ 5 points de pourcentage au-dessus de la moyenne nationale.

Les résultats indiquent aussi des taux de sous-dénombrement élevés pour d'autres groupes ethniques, notamment la population d'origine hispanique. De même, les taux de sous-

¹ Gregg Diffendal, Undercount Research Staff, Statistical Research Division, Bureau of the Census, Washington (D.C.) 20233, E.-U. Cet article est un compte rendu des recherches qui ont été faites par un membre du personnel du Census Bureau. Les opinions qui y sont exprimées sont celles de l'auteur et ne reflètent pas nécessairement la position du Census Bureau.

- BARR, A.J., GOODNIGHT, J.H., SALL, J.P., BLAIR, W.H., et CHILKO, D.M. (1979). SAS User's Guide. SAS Institute, Raleigh, North Carolina.
- BINDER, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-292.
- COCHRAN, W.G. (1977). *Sampling Techniques*. New York: John Wiley.
- DISKIN, B.A. (1985). Microcomputers in developing country statistical offices: current use and a look to the future. Actes de la 45^e assemblée de l'Institut international de statistique, Amsterdam.
- FRANCISCO, C.A. (1987). Estimation of quantiles and the interquartile range in complex surveys. Thèse de doctorat non publiée, Iowa State University, Ames, Iowa.
- FRANKEL, M.R. (1971). *Inference from Survey Samples*. Institute of Social Research, University of Michigan, Ann Arbor.
- FULLER, W.A. (1975). Regression analysis for sample survey. *Sankhyā C* 37, 117-132.
- FULLER, W.A., et SULLIVAN, G. (1987). Gamma Post Stratification. Report to the U.S. Bureau of the Census. Département de statistique, Iowa State University, Ames, (Iowa).
- HERRAMAN, C. (1968). Sums of squares and products matrix. *Applied Statistics*, 17, 289-292.
- HIDIROGLOU, M.A. (1974). Estimation of regression parameters for finite populations. Thèse de doctorat non publiée, Iowa State University.
- HIDIROGLOU, M.A., FULLER, W.A., et HICKMAN, R.D. (1980). *SUPER CARP*, Département de statistique, Iowa State University, Ames, Iowa.
- KISH, L. (1965). *Survey Sampling*. New York: John Wiley.
- LONGLEY, J.W. (1967). An appraisal of least squares programs for the electronic computer from the point of view of the user. *Journal of the American Statistical Association*, 62, 819-841.
- PARK, H.J. (1987). Univariate Analysis in PC CARP. Creative Component pour le M.S. non publiée, Iowa State University, Ames, Iowa.
- RAO, J.N.K., et SCOTT, A.J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *Annals of Statistics* 12, 46-60.
- WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

BIBLIOGRAPHIE

Figure 2. Résultats de l'analyse unidimensionnelle pour les ménages non agricoles (Exemple D).

ANALYSE UNIDIMENSIONNELLE 1

Variable de classification: caractère agricole état: 1

Nombre d'éléments de l'échantillon dans la sous-population = 111

Variable dépendante: âge

Variance de la sous-population = 4.25645D + 02

C.V. de la sous-population = 7.14101D-01

Moyenne de la sous-population

Estimation	E.T.	C.V.	DFFF
2.8891089D + 01	2.2543875D + 00	7.80305D-02	9.86336D-01

Valeurs extrêmes des éléments de l'échantillon dans la sous-population

Valeurs les plus faibles	Nombre de valeurs observées	strate	grappe	observation	poids
1.000D + 00	1	32	1	1	3.000D + 00
2.000D + 00	1	15	1	1	2.000D + 00
3.000D + 00	3	10	2	2	3.000D + 00

Identification de la première

Valeurs les plus élevées	Nombre de valeurs observées	strate	grappe	observation	poids
7.400D + 01	2	29	1	1	3.000D + 00
7.100D + 01	2	7	1	1	2.000D + 00
7.000D + 01	4	7	1	1	2.000D + 00

Quantiles

Estimation	E.T.	Intervalle de confiance de 95%
2.2690811D + 00	7.3167585D - 01	(8.05729D - 01, 3.73243D + 00)
4.2364814D + 00	1.1977759D + 00	(3.34942D + 00, 8.14058D + 00)
7.7750203D + 00	1.3563759D + 00	(5.36691D + 00, 1.07924D + 01)
1.3652930D + 01	1.4225576D + 00	(9.99238D + 00, 1.56826D + 01)
1.9449315D + 01	2.2740912D + 00	(1.57192D + 01, 2.48156D + 01)
4.5698071D + 01	4.7577709D + 00	(3.58936D + 01, 5.49247D + 01)
6.2787426D + 01	2.3472775D + 00	(5.51417D + 01, 6.45308D + 01)
6.5923423D + 01	1.22228344D + 00	(6.43837D + 01, 6.92750D + 01)
7.1714993D + 01	1.1425033D + 00	(7.05401D + 01, 7.40000D + 01)

Ecart interquartile

Estimation 3.2045141D + 01
E.T. 4.2434890D + 00
grand nombre de paramètres statistiques ont été calculés pour les analyses B et C que pour l'analyse A. Quant à l'analyse D, elle a nécessité quatre lectures de données, soit deux pour chaque analyse à une variable.

REMERCIEMENTS

Outre le U.S. Census Bureau, le Soil Conservation Service du Département de l'Agriculture de E.-U. a apporté son soutien au projet PC CARP. Nous tenons à remercier les arbitres et les éditeurs pour leurs commentaires utiles. Les personnes dans les pays en voie de développement qui s'intéressent au programme PC CARP sont priées de s'adresser à: PC CARP, International Statistical Programs Center, U.S. Bureau of the Census, Washington (D.C.), E.-U. 20233. Les autres personnes sont priées de s'adresser à: PC CARP, Statistical Laboratory, Iowa State University, Ames (Iowa), E.-U. 50011. Des exemplaires du programme ont été offerts à des organismes répartis dans plus de 20 pays.

Figure 1. Résultats de l'analyse C; âge moyen selon le sexe et l'origine raciale.

MOYENNES DE SOUS-POPULATION

Variable dépendante: âge							
Catégorie		Estimation		E.T.		C.V.	
						DEFF	
Sexe = 1.0000		Origine raciale = 1.0000		3.0681D + 01		2.0197D - 02	
Sexe = 1.0000		Origine raciale = 2.0000		3.13016D + 01		2.5193D - 02	
Sexe = 1.0000		Origine raciale = 3.0000		7.88580D - 01		9.5384D - 01	
Sexe = 1.0000		Origine raciale = 4.1579D + 01		2.2411D + 00		6.5610D - 02	
Sexe = 2.0000		Origine raciale = 1.0000		1.33742D + 01		2.3845D - 02	
Sexe = 2.0000		Origine raciale = 2.0000		3.18904D - 01		1.2588D + 00	
Sexe = 2.0000		Origine raciale = 3.0000		9.53816D - 01		5.5468D - 02	
Sexe = 2.0000		Origine raciale = 5.0000		1.71957D + 01		1.1389D + 00	

Une variable qui prend la valeur 1 pour chaque observation (variable constante) a été créée par le programme. Les analyses exécutées sont les suivantes:

- A. Revenu moyen de la population échantillonnée
- B. Revenu moyen par strate
- C. Âge moyen pour la classification selon le sexe et l'origine raciale
- D. Fonctions de distribution de l'échantillon selon l'âge pour les groupes agricoles et non agricoles.

L'analyse A (estimation du revenu moyen) a été exécutée au moyen de l'option "analyse de rapports", où le revenu servait de numérateur et la variable constante de dénominateur. Les estimations du revenu moyen par strate (analyse B) ont été calculées directement au moyen de l'option "moyenne de strate". L'analyse C a été exécutée au moyen de l'option "moyenne de sous-population" en croisant les variables de classification "sexe" et "origine raciale" et en définissant l' "âge" comme la variable dépendante. Les résultats de l'analyse C sont reproduits dans la figure 1. Les symboles "*****" qui paraissent sous la classification "sexe = 2 origine raciale = 2" indiquent qu'il n'y avait pas d'observations relatives à cette catégorie. Les valeurs de l'effet du plan mettent en relief l'importance de tenir compte du plan de sondage dans l'estimation de la variance. Par exemple, l'effet du plan pour l'estimation correspondant à la catégorie "sexe = 1 et origine raciale = 3" est environ 2. Cela signifie que la variance estimée de la moyenne de l'échantillon pour un échantillonnage aléatoire simple équivalait à la moitié de la variance estimée pour un échantillonnage en grappes stratifié. Les caractéristiques de la distribution de l'âge pour chacun des deux états de la variable "caractère agricole" ont été estimées au moyen de l'option "analyse à une variable". La figure 2 donne les résultats de l'analyse à une variable pour les ménages non agricoles. Toutes les variances et les estimations d'erreurs types indiquées dans cette figure tiennent compte du plan de sondage.

La durée d'exécution des analyses A, B, C et D pour les 2,400 observations a été de 70, 135, 120 et 360 secondes respectivement. L'exécution s'est faite sur IBM PC AT, les données étant stockées sur le disque dur et lues en structure non imposée. Les taux de sondage des strates n'ont pas été introduits dans le programme. Les résultats ont été acheminés au récepteur de contrôle et à un fichier-disque. Les effets du plan pour les estimations ont dû être considérés dans les quatre analyses. Les trois premières analyses ont exigé chacune une seule lecture des données. Un plus

4.6 Regroupement de strates

Aux fins du calcul de la variance, l'utilisateur peut recourir à cette option afin d'éliminer les strates à une grappe. Lorsque cette option est appliquée, chaque strate à une grappe est assimilée à la strate suivante dans la série de données. On modifie en conséquence le numéro de strate et le numéro de grappe des enregistrements touchés par cette fusion. S'il existe des taux de sondage de strates, les nouveaux taux sont définis par

$$f_i^* = (n_i f_{i-1}^* + n_{i+1} f_{i+1}^* - 1) / (n_i + n_{i+1}),$$

où la strate i , avec $n_i = 1$, a été combinée avec la strate $i + 1$. Ces nouveaux taux sont enregistrés dans un fichier annexe pour usage ultérieur. L'ordre de classement des strates influera sur le mode de regroupement des séries de données et les taux de sondage de strates. Le programme exige une lecture de données additionnelle lorsque l'utilisateur recourt à l'option "regroupement de strates" ou à l'option "à deux degrés".

4.7 Imputation par la méthode "hot deck"

PC CARP exige des séries de données complètes pour l'analyse. Pour le contrôle et l'imputation des données, beaucoup d'utilisateurs choisiront d'écrire un programme spécial ou encore utiliseront l'un des programmes offerts à cette fin dans la gamme PC.

Un programme d'imputation par "hot deck", appelé PRE CARP, peut être fourni avec le PC CARP pour ceux qui en font la demande. La méthode "hot deck" consiste à remplacer une valeur manquante par la valeur de l'article correspondant de l'enregistrement qui vient immédiatement avant dans le fichier de données. PRE CARP permet à l'utilisateur de définir une variable de classification qui peut renfermer jusqu'à 10 catégories de telle sorte que la valeur manquante est remplacée par la valeur de l'enregistrement précédent dans la même catégorie. PRE CARP créera aussi une variable indicatrice pour chaque variable pour laquelle il y a des valeurs manquantes. Cette variable indicatrice peut être ensuite utilisée avec l'option "analyse de sous-population" pour calculer des moyennes en fonction des observations originales.

5. EXEMPLES

Dans cette section, nous effectuons plusieurs analyses à l'aide d'une série de données composées et présentons les durées d'exécution. Le but des passages d'essai n'est pas d'étudier toutes les combinaisons possibles de facteurs qui peuvent influencer la durée de l'exécution mais d'estimer le temps requis pour exécuter quelques-unes des analyses offertes dans le programme. Les données d'essai ont été constituées à partir d'un sous-ensemble de la seconde édition de la National Health and Nutrition Examination Survey (NHANES II). La série ainsi constituée comprend 2400 observations réparties en 32 strates. Chaque strate compte deux unités primaires d'échantillonnage et ces unités d'échantillonnage sont de taille variée. De plus, chaque observation est affectée d'un poids de sondage non nul.

Les variables contenues dans la série de données sont les suivantes:

- 1. Sexe
1 = homme, 2 = femme
- 2. Origine raciale
1 = blanc, 2 = noir, 3 = autre
- 3. Caractère agricole
1 = ménage non agricole, 2 = ménage agricole
- 4. Revenu
revenu du ménage (en milliers de dollars)
- 5. Âge
âge (en années).

où X est la matrice $n \times p$ des valeurs de la variable indépendante, Y est le vecteur à n dimensions des valeurs de la variable dépendante, W est une matrice dont la diagonale est formée des poids des observations et le reste est formé de zéros, et n est le nombre total des observations.

La variance de b est estimée par

$$V(b) = (X'WX)^{-1}G^w(X'WX)^{-1}.$$

La matrice G^w est

$$G^w = C \sum_{l=1}^L \sum_{n_l} h_l (d_{ij} - \bar{d}_{i.})(d_{ij} - \bar{d}_{i.}),$$

où

$$\bar{d}_{ij} = \sum_{m_{jk}}^{n_l} X_{ijk} \hat{v}_{ijk} W_{ijk}$$

$$\hat{v}_{ijk} = Y_{ijk} - b'X_{ijk}$$

$$\bar{d}_{i.} = n_i^{-1} \sum_{n_l}^{j=1} \hat{d}_{ij}$$

en ce que la matrice G^w est substituée à $(X'WX)^2$.
Le programme calcule une fonction des observations R^2 multiple pour les modèles qui ont une ordonnée ou une abscisse à l'origine. De plus, il effectue toujours un test F pour la régression générale et offre à l'utilisateur la possibilité de tester des sous-ensembles de coefficients.

4.4 Régression logistique

Cette option permet d'obtenir des estimations du modèle logistique à plusieurs variables. Les algorithmes destinés à la régression logistique ont été élaborés après que la version initiale de PC CARP a vu le jour. Comme la fonction de moyenne pour le modèle logistique est non linéaire dans les paramètres, les estimateurs sont calculés à l'aide d'un algorithme d'itération fondé sur les moindres carrés pondérés. Les variances des estimateurs sont calculées au moyen des méthodes décrites dans Fuller (1975) et appliquées à l'estimation non linéaire. Voir aussi Binder (1983). Au point de vue des modalités d'application, l'option "régression logistique" est en tout point semblable à l'option "régression". Par exemple, les variables dépendantes et indépendantes y sont définies de la même façon.

4.5 Stratification a posteriori

Après que l'on a élaboré la version originale du programme PC CARP, une option "stratification a posteriori" a été mise au point pour un bon nombre des estimateurs. La stratification a posteriori est censée être une opération par laquelle on corrige les poids de population connus des estimations pour certaines catégories, qui concordent avec des totaux de population connus. C'est ce que Fuller et Sullivan (1987) appellent la stratification a posteriori gamma. Le programme calcule la variance de l'estimateur de stratification a posteriori à partir d'une formule dans laquelle l'estimateur est exprimé comme la somme d'estimateurs de rapports.

où V_{YY} est la matrice estimée des covariances du vecteur des fréquences de case Y , I_{RC} est la matrice unité de dimension $RC \times RC$ et J_{RC} est un vecteur colonne de dimension $RC \times 1$, constitué de uns.

La matrice V_{pp} sert à calculer la variable à tester pour l'hypothèse de la proportionnalité. L'hypothèse nulle, en l'occurrence, est que les éléments intérieurs du tableau de la population sont le produit des proportions marginales. Voir Rao et Scott (1984) pour une analyse des tests relatifs à ce genre d'hypothèses. Le test dans PC CARP est fondé sur une approximation de Satterthwaite de la distribution du critère chi-carré de Pearson, construite comme s'il s'agissait de proportions multinomiales. L'approximation est valide pour n'importe quelle variable d'analyse.

4.2 Estimation de quantiles

L'option "analyse unidimensionnelle" permet notamment de calculer des estimations de quantiles ainsi qu'un estimateur de l'erreur type des quantiles. Lorsqu'on veut calculer des quantiles, il faut tout d'abord construire un estimateur de la fonction de distribution cumulative. Une première lecture des données permet de déterminer l'intervalle des observations de même que la moyenne et l'écart type de l'échantillon. Par la même occasion, on relève les trois plus grandes et les trois plus petites observations.

La fonction de distribution cumulative estimée est définie par

$$F_Y(x) = \left(\sum_m^t w_t Z_{St} \right)^{-1} \sum_m^t w_t Z_{St} I_Y(x),$$

où la sommation est étendue aux m éléments de l'échantillon, w_t est le poids d'échantillon, Z_{St} est une fonction indicatrice qui prend la valeur 1 si l'observation est comprise dans la sous-population d'intérêt et la valeur 0 dans le cas contraire, et $I_Y(x)$ est égale à 1 si $X < x$ et est égale à zéro dans le cas contraire. Le domaine de la variable est divisé en 100 intervalles et la fonction de distribution cumulative est estimée aux 101 points définis par cette division.

Le programme estime ensuite la matrice des covariances pour la fonction de distribution estimée évaluée à 25 points, $j = 1, 5, \dots, 96$. Les erreurs types estimées sont lissées à l'aide d'une moyenne mobile de trois points et une interpolation permet d'établir une erreur type estimée pour chacun des 101 points de la fonction de distribution estimée. Une interpolation linéaire permet de créer une fonction de distribution estimée qui est croissante au sens large. À l'aide des erreurs types lissées, le programme détermine une borne supérieure monotone croissante et une borne inférieure monotone croissante, qui délimitent un intervalle de confiance ponctuel de 95% pour la fonction de distribution. Ces bornes sont ensuite inversées pour former des intervalles de confiance de 95% pour les quantiles. L'écart interquartile et l'erreur type correspondante sont aussi estimés.

L'estimation de quantiles repose sur une théorie qui suppose l'existence d'une fonction de distribution de superpopulation sous-jacente avec une densité positive. Voir Francisco (1987) pour les aspects théoriques et Park (1987) pour les aspects relatifs aux calculs.

4.3 Estimation par régression

Les estimations des coefficients d'un modèle de régression linéaire sont calculées par la méthode des moindres carrés pondérés. Suivant la méthode décrite dans Fuller (1975), le programme calcule un estimateur de la matrice des covariances du vecteur des coefficients en tenant compte du plan de sondage.

Le vecteur des coefficients est estimé par

$$b = (X'WX)^{-1}X'WY,$$

lecture de données exige beaucoup d'espace mémoire. Cependant, lorsqu'on travaille avec de grands échantillons, la suppression de lectures entières compense l'utilisation d'espace mémoire additionnel. Le programme exécute régulièrement des vérifications visant à déceler des erreurs de calcul comme un diviseur nul. Si, par exemple, l'utilisateur introduit une série de données avec une strate à une seule grappe, le programme attribuera une variance nulle à la strate, complètera les calculs et fera imprimer un message d'erreur qui signalera la strate en question.

Le système de traitement d'erreurs a été conçu de manière à prévenir les arrêts de programme causés par des erreurs de l'utilisateur qui peuvent être corrigées facilement. Le programme renferme des fonctions de vérification pour l'absence de réponse, l'inexactitude de noms de fichiers et l'inexactitude de définitions de variables d'analyse. Si une erreur est décelée, le PC CARP permet à l'utilisateur de réintroduire les données ou de sortir du programme.

On a évalué la précision du programme en construisant des exemples d'application puis en comparant les résultats à ceux obtenus à l'aide du programme principal SUPER CARP. La série de données de Longley (1967) a servi à évaluer la précision du programme de régression. Des vérifications supplémentaires ont été faites à l'aide de PROC MATRIX du logiciel SAS. Voir Barr et coll. (1979). On a constaté que PC CARP avait un niveau de précision numérique équivalent à celui des modèles principaux. On a aussi vérifié la cohérence interne de PC CARP en calculant des estimateurs équivalents à l'aide d'options différentes (p. ex., calcul d'une moyenne de sous-population au moyen de l'option "analyse de sous-population" et de l'option "analyse de rapports").

Lorsque PC CARP demande de l'information à l'utilisateur, une série de questions à réponse brève apparaît à l'écran avec des instructions détaillées. La première série d'affichages demande à l'utilisateur de fournir des renseignements sur l'organisation et l'emplacement des données. Des options "Help" et "Go Back" peuvent être utilisées à plusieurs endroits.

La seconde phase de l'exécution du programme a trait à la définition de l'analyse. Dans cette phase, l'utilisateur choisit le genre d'analyse, les options pertinentes et les variables d'analyse. Un nombre indéfini d'analyses peuvent être exécutées à l'aide des données définies dans la phase 1.

4. CARACTÉRISTIQUES SPÉCIALES

4.1 Tableau à double entrée

Comme nous l'avons vu dans la Section 2, cette option fournit automatiquement à l'utilisateur quatre tableaux dont les éléments sont déterminés par le genre de liaisons qui ont été définies dans la construction du tableau à propos des fréquences par case.

Nous exposons ci-dessous la méthode qui a servi à construire le tableau des proportions de case et la matrice estimée des covariances des proportions. Supposons que le tableau est formé de R lignes et de C colonnes et posons Y_{rc} comme la fréquence estimée pour la case rc . Soit Y le vecteur colonne $RC \times 1$ des fréquences de case, créé en reproduisant l'une au-dessous de l'autre, dans l'ordre normal, les colonnes du tableau $R \times C$. Soit

$$Y_{..} = \sum_{R} \sum_{C=1}^C Y_{rc},$$

$$P_{rc} = Y_{..}^{-1} Y_{rc}$$

le total estimé de la population et la proportion de case estimée pour la case rc respectivement. Soit P le vecteur colonne $RC \times 1$, analogue à Y , composé des RC valeurs P_{rc} disposées en colonnes. La matrice estimée des covariances pour P est

$$V_{PP} = Y_{..}^{-2} [I_{RC} - (P \otimes J_{RC})] V_{YY} [I_{RC} - (P \otimes J_{RC})]',$$

élémentaires a permis de réduire au minimum la fréquence de lecture des fichiers. Une seule lecture des données suffit pour calculer la plupart des estimateurs et leur variance.

Le programme peut calculer des estimateurs pour la population totale, pour chaque strate ou pour des sous-populations déterminées. Ces estimateurs sont pour la plupart des fonctions de totaux d'échantillon pondérés. Si l'on veut calculer, par exemple, l'estimateur des rapports $R_1 = Y_1/X_1$ et $R_2 = Y_2/X_2$, on recueille les totaux pour Y_1, X_1, Y_2 , et X_2 . Si l'estimation concerne la population totale, on forme ces totaux par une seule lecture des données. Dans le cas des estimations de strate, les totaux peuvent être accumulés, et combinés si nécessaire, puis restitues strate par strate. Comme les données sont groupées par strate, une lecture suffit pour déterminer les totaux de strate pour n'importe quelle quantité de strates. Pour ce qui a trait aux estimations relatives à une sous-population, une seule lecture des données pourrait ne pas suffire si le nombre de catégories définies par la structure de classification est élevé. L'analyse de régression et l'analyse unidimensionnelle exigent deux lectures de données.

Tous les estimateurs, à l'exception des estimateurs d'agrégats, sont des fonctions non linéaires des moments pondérés de l'échantillon. Il est donc nécessaire d'utiliser une méthode conforme à une fonction non linéaire pour estimer la variance de la distribution approximative de ce genre d'estimateurs. Voir Wolter (1985) pour une analyse de l'estimation de la variance pour des enquêtes complexes. La méthode de Taylor (méthode des écarts statistiques) est la méthode d'estimation de la variance utilisée dans PC CARP. Il a été démontré que la méthode de Taylor était en règle générale aussi efficace sinon plus efficace que les autres méthodes d'estimation de la variance des paramètres statistiques à l'étude (p. ex.: les rapports). Voir, par exemple, Frankel (1971). La méthode de Taylor pour l'estimateur de rapports est décrite dans des ouvrages aussi courants que celui de Cochran (1977) et la même méthode pour l'estimateur de coefficients de régression est décrite dans Fuller (1975).

Dans la plupart des cas, l'estimateur et sa variance estimée peuvent être calculés par la même lecture de données parce que l'approximation de Taylor du premier ordre de la variance peut être exprimée en fonction des variances des totaux. Par exemple, l'approximation de Taylor du premier ordre de $R = Y/X$ est

$$\hat{R} \approx R + X^{-1} (Y - RX),$$

où $R = Y/X$ est le rapport des totaux réels. Par conséquent, la variance estimée d'un rapport $R = Y/X$ peut être calculée à l'aide de la variance estimée des totaux de Y, X , et $(Y - X)$. De même, la matrice estimée des covariances pour $R_1 = Y_1/X_1$ et $R_2 = Y_2/X_2$, peut être calculée à l'aide des variances estimées des totaux de $Y_1, X_1, (Y_1 - X_1), Y_2, X_2, (Y_2 - X_2), (Y_1 - Y_2), (Y_1 - X_2), (Y_2 - X_1)$, et $(X_1 - X_2)$.

L'algorithme servant au calcul de la moyenne pondérée, des sommes des carrés pondérées et des matrices de produits croisés est décrit dans Herraman (1968). Pour les valeurs d'échantillon X_i et les poids correspondants $\{W_i\}$, la suite des moyennes pondérées \bar{X}_K , et des sommes des carrés corrigées pondérées S_K est calculée à l'aide des formules suivantes

$$\bar{X}_K = \bar{X}_{K-1} + a_K d_K \text{ et } S_K = S_{K-1} + D_K - D_K a_K,$$

$$\text{où } d_K = X_K - \bar{X}_{K-1}, a_K = W_K (\sum_{i=1}^K W_i)^{-1}, \text{ et } D_K = d_K^2 W_K.$$

Il est possible de calculer simultanément trois composantes de variance pour n'importe quel estimateur donné. Il s'agit de la composante de variance du premier degré, de la composante de variance optionnelle du second degré et de la variance optionnelle d'un échantillonnage aléatoire simple utilisée dans le calcul de l'effet du plan. Calculer toutes ces composantes par une seule

pour l'estimation de la fonction logistique et la stratification a posteriori. Ces fonctions supplémentaires sont décrites dans les sections 4.4 et 4.5. Un "X" sous la rubrique "matrice des covariances" signifie que l'on peut calculer la matrice des covariances d'un vecteur d'estimations du genre de celles indiquées dans la colonne de gauche. L'erreur type peut être calculée pour tous les paramètres statistiques mais la matrice des covariances d'un vecteur ne peut l'être que pour un nombre limité de paramètres. En outre, le coefficient de variation peut être calculé pour un grand nombre de paramètres. L'effet du plan, désigné par DFF, peut être calculé sur demande pour un grand nombre de paramètres. Voir Kish (1965) pour une description de l'effet du plan. Les analyses de population (agrégat, rapport et différence de rapports) et les analyses de strate s'effectuent d'une manière simple. Pour ce qui est des autres analyses (analyse de sous-population, de tableau à double entrée, de régression et analyse unidimensionnelle), on trouvera les détails pertinents à la section V.

Les analyses de sous-population offrent à l'utilisateur la possibilité de croiser des variables de classification. Grâce à cela, l'utilisateur peut créer de nouvelles structures de classification à l'aide de deux variables de classification ou plus. Supposons, par exemple, que les données d'entrée comprennent les variables de classification "âge" (6 niveaux), "sexe" (2 niveaux) et "niveau d'instruction" (5 niveaux). Si nous combinons ces trois variables, nous allons obtenir une nouvelle structure de classification à soixante niveaux. Suivant cette structure, l'utilisateur peut établir des estimations pour un nombre indéterminé de variables dépendantes. L'analyse des tableaux à double entrée est définie par deux variables de classification et une variable dépendante. Il est possible de définir plus d'une variable dépendante pour une paire de variables de classification. Pour chaque variable dépendante, le programme calcule les tables de fréquences de cases ainsi que les tables de proportions fondées sur les totaux de ligne, les totaux de colonne et le grand total. Il calcule aussi des erreurs types pour tous les estimateurs et produit une fonction des observations visant à tester l'hypothèse de la proportionnalité. Cette fonction est fondée sur une approximation de Satterthwaite de la distribution du critère chi-carré de Pearson. Voir aussi Rao et Scott (1984).

L'analyse de régression par les moindres carrés pondérés permet de calculer des estimations de coefficients ainsi qu'une matrice estimée des variances-covariances qui tient compte du plan de sondage. Ces calculs sont présentés dans Fuller (1975) et repris en bref dans Hidiroglou et coll. (1980). Des tests F à plusieurs degrés de liberté pour des séries de coefficients et les variables et des valeurs estimatives.

L'analyse unidimensionnelle produit des statistiques qui décrivent la distribution d'une variable. L'utilisateur définit la variable d'intérêt et détermine une sous-population en choisissant une catégorie d'une variable de classification. Ainsi, l'utilisateur pourrait choisir de recueillir des données sur le revenu personnel des membres de la catégorie des professions libérales dans la classification des professions. Le programme produit en l'occurrence des estimations de la moyenne de sous-population, de la variance, de la fonction de distribution, de quantiles et de l'écart interquartile.

3. DÉTAILS SUR LE PROGRAMME

PC CARP est écrit presque entièrement en FORTRAN, le langage de programmation scientifique le plus connu, et c'est le compilateur Professional FORTRAN d'IBM qui a été choisi pour le projet. Une petite partie du code — quelques sections de l'interface avec l'utilisateur — est rédigée en langage assembleur d'IBM.

Durant l'élaboration du programme, on a surtout cherché à créer une interface conviviale et à réduire au minimum le nombre de fois que les données doivent être parcourues. On a réussi à créer une interface conviviale en mettant en place un système de réponse interactif à écran. Par ailleurs, un algorithme monopasse pour l'estimation de la variance de paramètres statistiques

PC CARP traite avec autant de facilité et d'efficacité les longues et les petites séries de données. Il ne pose pas de limite pour le nombre de strates ou de grappes qui peuvent être incluses dans une série de données et peut accepter jusqu'à 50 paramètres à la fois. Il accepte, par ailleurs, les fichiers-disques à format de bloc fixe ou à structure libre.

Le programme peut servir à calculer les variances pour des échantillons à un ou à deux degrés avec correction pour population finie. Dans le cas des échantillons à plus de deux degrés, les termes de correction pour population finie n'existent qu'à deux niveaux. En ce qui concerne les échantillons à deux degrés, le programme calcule les taux de sondage à l'intérieur des grappes à partir des taux de sondage des strates et des poids des observations.

En règle générale, on connaît la strate et la grappe (unité primaire d'échantillonnage) auxquelles appartient chaque observation du fichier et on connaît aussi le poids de ces observations, qui correspond à l'inverse de la probabilité de sélection. L'utilisateur est libre d'introduire les taux de sondage du premier degré. Les plans de sondage simples, comme l'échantillonnage aléatoire simple, n'exigent pas tous ces renseignements. Dans de tels cas, il est possible d'avoir moins de paramètres.

Dans le cas d'un échantillonnage stratifié, toutes les observations qui sont contenues dans la même strate doivent être regroupées. S'il s'agit d'un échantillonnage par grappes, toutes les observations contenues dans la même grappe doivent être regroupées.

Le tableau 1 contient une liste des paramètres statistiques qui peuvent être calculés à l'aide du PC CARP. Outre les fonctions énumérées dans le tableau 1, il existe des cartes supplémentaires

Tableau 1

Fonctions d'analyse disponibles dans PC CARP

Analyse	Coef- ficient de variation	Matrice des co- variances	Effet du plan	Remarques
<i>Analyses de populations</i>				
Estimation d'agrégats	X	X	X	50 variables au maximum
Estimation de rapports	X	X	X	50 variables au maximum sans les covariances, 15 avec covariances
Différence de rapports			X	15 variables au maximum
<i>Analyses de strates</i>				
Aggrégats	X		X	50 variables au maximum
Moyennes	X		X	50 variables au maximum
Proportions	X		X	50 variables au maximum
<i>Analyses de sous-populations</i>				
Aggrégats	X		X	Classement recoupé Variables multiples
Moyennes	X		X	Classement recoupé Variables multiples
Proportions	X		X	Classement recoupé Variables multiples
Rapports	X		X	Classement recoupé Variables multiples
<i>Autres analyses</i>				
Tableau à double entrée	X			50 cas au maximum, test de proportionnalité
Régression		X		50 variables au maximum Tests à plusieurs degrés de liberté, valeurs prédites résidus
Analyse unidimensionnelle			X	Variables multiples, fonctions de distribution cumulatives empiriques, quantiles

Logiciel d'ordinateur personnel pour l'estimation de la variance dans des enquêtes complexes

DAN SCHNELL, WILLIAM J. KENNEDY, GARY SULLIVAN,
HEON JIN PARK et WAYNE A. FULLER¹

RÉSUMÉ

Le présent article contient la description d'un programme d'ordinateur personnel servant à l'estimation de la variance pour de grandes enquêtes. Ce programme, connu sous le nom de PC CARP, permet de calculer des estimations pour des agrégats, des rapports, des moyennes, des quantiles et des coefficients de régression et d'estimer les variances correspondantes.

MOTS CLÉS: Échantillon d'enquête; estimation de la variance; logiciel d'enquête.

1. INTRODUCTION

L'analyse de données d'enquête comporte habituellement un grand nombre d'observations et des calculs relativement complexes en ce qui a trait à la variance. Grâce aux derniers perfectionnements qui leur ont été appliqués, les ordinateurs personnels peuvent maintenant servir à traiter des données d'enquêtes complexes. Nous nous proposons dans cet article de décrire un programme d'ordinateur personnel qui est destiné à l'analyse de données d'enquête et qui a été mis au point à l'université Iowa State.

La création d'un logiciel d'ordinateur personnel pour l'estimation de la variance a été le résultat de la collaboration entre l'université Iowa State et le International Statistical Programs Center du U.S. Census Bureau. Par ce projet, le Census Bureau visait à offrir aux pays en voie de développement un logiciel qui pourrait être utilisé sur place pour traiter des données recueillies sur place. Le projet de l'université Iowa State s'inscrivait dans un programme du Census Bureau qui avait pour but de créer des logiciels pour la gestion des enquêtes et le contrôle et la totalisation des données.

Au début des années 1970, l'université Iowa State, se fondant sur les travaux d'Hidiroglou (1974) et Fuller (1975), élaborait un programme ayant pour but de calculer des coefficients de régression et la matrice estimée des covariances de ces coefficients pour des données d'enquête. Ce programme, connu sous le nom de SUPER CARP, devait plus tard être modifié de manière à inclure l'estimation d'agrégats et de rapports, les statistiques de sous-population, les tableaux à double entrée et les échantillons à deux degrés. La dernière révision de SUPER CARP remonte à 1980. Ce programme a ouvert la voie à la création de logiciels pour ordinateurs personnels.

Pour cette raison, le programme d'ordinateur personnel a été baptisé PC CARP.

2. CARACTÉRISTIQUES DU PROGRAMME

PC CARP est conçu pour les modèles IBM PC, IBM PC/XT, IBM PC/AT et les ordinateurs compatibles. Il exige une mémoire d'au moins 410 kilo-octets ainsi qu'un coprocesseur "mathématique".

¹ Dan Schnell, Centers for Disease Control, 1600 Clifton Road, NE, Atlanta, (Géorgie) 30333. William J. Kennedy, Gary Sullivan, Heon Jin Park et Wayne A. Fuller, Département de statistique, Iowa State University, Ames, Iowa 50011, États-Unis.

- U.S. Bureau of the Census (1986). Additional results from the SIPP telephone test. Note de service de J. Coder à A. Norton, avril 9, 1986.
- U.S. Bureau of the Census (1985). SIPP research on SIPP oversampling/subsampling. Note de service de R. Singh à G. Shapiro et D. Kasprzyk, août 12, 1985.
- WEIDMAN, L. (1987). Examination of relationships between actual and reported changes in the SIPP. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 216-220.
- WEIDMAN, L. (1986). Investigation of gross changes in income reciprocity from the Survey of Income and Program Participation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 231-236.
- WHITE, G., et HUANG, H. (1982). Mover follow-up costs for the Income Survey Development Program. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 376-381.
- YCAS, M., et LININGER, C. (1981). The Income Survey Development Program: Design features and initial findings. *Social Security Bulletin*. Baltimore, Md.: Social Security Administration, 44.

- KALTON, G. (1983). *Compensating for missing survey data*. Survey Research Center, University of Michigan.
- KALTON, G., et KASPRZYK, D. (1986). Le traitement des données d'enquête manquantes. *Techniques d'enquête*, 12, 1-16.
- KALTON, G., et KASPRZYK, D. (1982). Imputing for missing survey responses. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 22-31.
- KALTON, G., KASPRZYK, D., et McMILLEN, D.B. (1988). Nonsampling errors in panel surveys. In *Panel Surveys* (éds. D. Kasprzyk, G. Duncan, et M.P. Singh), New York: John Wiley, à paraître.
- KALTON, G., et LEPKOWSKI, J. (1985). Following rules in SIPP. *Journal of Economic and Social Measurement*, 13, 319-329.
- KALTON, G., LEPKOWSKI, J., et LIN, T. (1985). Compensating for wave nonresponse in the 1979 ISDP research panel. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 372-377.
- KALTON, G., McMILLEN, D., et KASPRZYK, D. (1986). A review of nonsampling error issues in the Survey of Income and Program Participation. *Proceedings of the U.S. Bureau of the Census Second Annual Research Conference*, 147-165.
- KALTON, G., et MILLER, M. (1986). Effects of adjustments for wave nonresponse on panel survey estimates. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 194-199.
- KASPRZYK, D. (1983). Social Security number reporting, the use of administrative records, and the multiple frame design in the Income Survey Development Program. Dans *Technical, Conceptual, and Administrative Lessons of the Income Survey Development Program* (ISDP), (éd. M. David), Washington, D.C.: Social Science Research Council, 171-198.
- KOBILARCIC, E., et SINGH, R. (1986). SIPP longitudinal estimation for persons' characteristics. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 214-219.
- LAMAS, E.J., et MCNEIL, J.M. (1987). An analysis of the SIPP asset and liability feedback experiment. *Proceedings of the Social Statistics Section, American Statistical Association*, 194-199.
- LEPKOWSKI, J. (1988). The treatment of wave nonresponse in panel surveys. Dans *Panel Surveys*, (éds. D. Kasprzyk, G. Duncan, et M.P. Singh), New York: John Wiley, à paraître.
- McMILLEN, D.B., et HERRIOT, R. (1985). Toward a longitudinal definition of households. *Journal of Economic and Social Measurement*, 13, 349-360.
- MOORE, J.C., et KASPRZYK, D. (1984). Month-to-month reciprocity turnover in the ISDP. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 726-731.
- MOORE, J.C., et MARQUIS, K. (1987). Utilisation des données de dossiers administratifs pour l'évaluation de la qualité des estimations d'enquêtes. Communication présentée au Symposium international sur l'utilisation statistique de données administratives, 1987, Ottawa, Canada.
- NELSON, D., McMILLEN, D., et KASPRZYK, D. (1985). An overview of the Survey of Income and Program Participation: Update I. *SIPP Working Paper Series No. 8401*. Washington, D.C.: U.S. Bureau of the Census.
- OLSON, J. (1980). *Reports from the Site Research Test*, (éd. J. Olson). Office of the Assistant Secretary for Planning and Evaluation, Department of Health and Human Services, United States.
- ROMAN, A.M., et O'BRIEN, D.V. (1984). The student follow-up investigation of the 1979 ISDP. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 732-737.

- CHAKRABARTY, R. (1986). Composite estimation for SIPP: A preliminary report. *SIPP Working Paper Series No. 8610*. Washington, D.C.: U.S. Bureau of the Census.
- CHAPMAN, D., BAILEY, L., et KASPRZYK, D. (1986). Méthodes de compensation de la non-réponse au U.S. Bureau of the Census. *Techniques d'enquête*, 12, 161-180.
- CITRO, C. (1985). Alternative definitions of longitudinal households in the Income Survey Development Program: Implications for annual statistics. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 381-386.
- CITRO, C., HERNANDEZ, D., et HERRIOT, R. (1986). Longitudinal household concepts in SIPP: Preliminary results. *Proceedings of the U.S. Bureau of the Census Second Annual Research Conference*, 598-619.
- CITRO, C., HERNANDEZ, D., et MOORMAN, J. (1986). Longitudinal household concepts in SIPP. *Proceedings of the Social Statistics Section, American Statistical Association*, Washington, D.C., 361-366.
- CODER, J. (1980). Some results from the 1979 Income Survey Development Program research panel. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 540-545.
- ERNST, L. (1988). Weighting issues for longitudinal household and family estimates. Dans *Panel Surveys* (eds. D. Kasprzyk, G. Duncan, G. Kalton, et M.P. Singh), New York: John Wiley, à paraître.
- ERNST, L., HUBBLE, D., et JUDKINS, D. (1984). Longitudinal family and household estimation in SIPP. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 682-687.
- FAY, R.E., et HUGGINS, V.J. (1988). Use of administrative data in SIPP longitudinal estimation. Article à présenter à la Section on Survey Research Methods, American Statistical Association, August 1988.
- GBUR, P., et DURANT, S. (1987). Testing telephone interviewing in the Survey of Income and Program Participation and some early results. Communication présentée au International Symposium on Telephone Survey Methodology, 1987, Charlotte, North Carolina.
- HEERINGA, S., et LEPKOWSKI, J. (1986). Longitudinal imputation for the SIPP. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 206-210.
- HILL, D. (1987). Response errors around the seam: Analysis of change in a panel with overlapping reference records. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 210-215.
- HUANG, H. (1984). Obtaining a cross-sectional estimate from a longitudinal survey: Experience of the ISDP. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 670-675.
- HUBBLE, D., et JUDKINS, D. (1986). Measuring the bias in gross flow estimates in the presence of auto-correlated response errors. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 237-242.
- JEAN, A., et McARTHUR, E. (1987.) Tracking Persons Over Time. *SIPP Working Paper Series No. 8701*, Washington, D.C.: U.S. Bureau of the Census.
- JEAN, A., et McARTHUR, E. (1984). Some data collection issues for panel surveys with application to the Survey of Income and Program Participation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 745-750.
- JUDKINS, D., HUBBLE, D., DORSCH, J.R., McMILLEN, D., et ERNST, L. (1984). Weighting of persons for SIPP longitudinal tabulations. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Washington, D.C., 676-681.
- KALTON, G. (1986). Handling wave nonresponse in panel surveys. *Journal of Official Statistics*, 2, 303-314.

Étude de vérification des enregistrements de la SIPP

Pour mettre en évidence la structure des erreurs relevées dans les résultats de la SIPP relative-
ment à la nature des prestations touchées et aux montants en cause, on peut, entre autres, réaliser
des études de validation des éléments communs à la fois aux données d'enquête et aux données
administratives. Une étude de ce genre a été amorcée dans le cadre du programme de la SIPP
dans le but d'en apprendre davantage sur la question de la qualité des réponses.
Cette étude devrait permettre de mieux juger de la qualité des données tirées de la SIPP et,
éventuellement, de produire des estimations précises des erreurs de réponse et de non-réponse,
afin que l'on puisse corriger les données d'enquête ou modifier les méthodes de collecte pour
améliorer la qualité de ces données. L'étude vise notamment à évaluer les aspects suivants:
1) la qualité des données déclarées par les répondants ayant indiqué qu'ils sont bénéficiaires
de divers programmes de transfert administrés par leur État ou le gouvernement fédéral; 2)
la qualité des données recueillies au sujet du montant des prestations versées en vertu de ces
programmes; 3) les corrélations démographiques pouvant être établies pour mesurer la qualité
des données; 4) l'importance des erreurs de classification; 5) l'incidence qu'a sur la qualité des
données le fait que les déclarations aient été faites par le répondant lui-même ou par un répon-
dant substitut; 6) les effets dus au renouvellement des bénéficiaires entre les cycles d'interview.
Quatre programmes à l'étude sont administrés par les États et les six autres, par le gouverne-
ment fédéral. Moore et Marquis (1987) présentent des résultats très provisoires dont il ressort
que les problèmes de déclaration sont différents pour l'aide aux familles à faible revenu avec
enfants à charge et pour les bons alimentaires: dans le cas de l'aide aux familles, il y a un pro-
blème certain de sous-déclaration ainsi qu'une difficulté à situer le versement des prestations
dans le temps, et dans le cas des bons alimentaires, seule la seconde difficulté a été observée.

6. CONCLUSION

Comme pour toutes les enquêtes permanentes de grande envergure, il faut poursuivre la
recherche si l'on veut mieux saisir l'incidence des techniques d'enquête sur la qualité des données
recueillies. Une enquête telle que la SIPP, dont la mise en oeuvre est complexe, exige que l'on
s'efforce d'essayer de comprendre le processus d'évaluation. Le large éventail de sujets soulevés
dans cet article – la collecte, l'estimation longitudinale, les définitions tenant compte du facteur
temps et l'erreur de réponse – renseignent le lecteur sur les questions qui ont retenu l'attention
tout au long du programme de mise au point de l'enquête et qui continuent de le faire depuis
que l'enquête elle-même a été lancée.

REMERCIEMENTS

L'auteur remercie sincèrement les reviseurs pour leurs commentaires utiles ainsi que Mme Hazel
Beaton pour les travaux de secrétariat qu'elle a assurés pendant la préparation de ce manuscrit.

BIBLIOGRAPHIE

- BOWIE, C., et KASPRZYK, D. (1987). A review of the use of administrative records in the Survey of
Income and Program Participation. *SIPP Working Paper Series No. 8721*, Washington, D.C.: U.S.
Bureau of the Census.
- BURKHEAD, D., et CODER, J. (1985). Gross changes in income reciprocity from the Survey of Income
and Program Participation. *Proceedings of the Section on Survey Research Methods, American
Statistical Association*, Washington, D.C., 351-356.

5. ERREUR DE RÉPONSE

L'erreur de réponse constitue un aspect d'un problème plus général, l'erreur non due à l'échantillonnage dont parlent Kalton, Kasprzyk et McMillen (1988). Les erreurs de réponse sont commises lorsque des données incorrectes sont inscrites dans les questionnaires, et ce pour diverses raisons notamment parce que le questionnaire est mal fait, la mémoire du répondant fait défaut ou le répondant n'est pas le bon. Pour permettre de mieux comprendre l'erreur de réponse en général, nous allons brièvement examiner ce problème dans le contexte des données de la SIPP sur les flux bruts et de l'étude de vérification des enregistrements.

Données de la SIPP sur les flux bruts

L'analyse des données relatives aux programmes effectuée mois par mois dans le cadre de l'ISDP a révélé que le renouvellement des bénéficiaires avait tendance à être déclaré plus souvent entre les cycles d'interview qu'au cours des cycles (Moore et Kasprzyk 1984). L'analyse des données tirées de la SIPP (Burkhead et Coder 1985) qui a porté sur la variation mensuelle du montant des prestations reçues pendant une période de douze mois a pour sa part accordé une attention particulière aux changements survenus entre le dernier mois d'une période de référence et le premier mois de la période de référence suivante.

Les résultats obtenus au moyen des données de l'ISDP et des données de la SIPP sont semblables, en ce sens que les changements observés sont irréguliers et sont clairement liés au plan d'interview. Les changements bruts sont nettement plus prononcés entre le dernier mois d'une période de référence et le premier mois de la période suivante. Hill (1987) s'est servi des données mensuelles recueillies au cours des cycles d'interviews de 1984 et de 1985 de la PSID (Panel Study of Income Dynamics/ étude par panel de la dynamique du revenu) pour mesurer les changements excessifs observés entre les cycles comparativement à ceux observés au cours des cycles et essayer d'en déterminer la cause. Il a trouvé que même si l'ordre des questions et la durée des période de référence étaient différents, les changements entre les cycles étaient plus nombreux que les changements pendant les cycles, comme pour la SIPP. Les causes principales du problème ne sont pas connues, mais la formulation des questions et la présentation du questionnaire de même que les défaillances de mémoire des répondants et l'interaction entre ces deux facteurs semblent être des causes probables.

Weidman (1986) a effectué une analyse empirique dans le but de déterminer s'il existait un lien évident entre certaines caractéristiques des répondants et les changements observés quant à un certain nombre de catégories de revenu perçu. Il n'en a trouvé aucun, sur une période de plusieurs mois consécutifs, entre les répartitions de changements bruts, le fait que les déclarations aient été faites en nom propre ou par personne interposée et neuf variables démographiques (l'âge, la race, le sexe, le niveau d'instruction, l'état matrimonial, la taille du ménage, le mode d'occupation du logement, le lien avec la personne de référence et la taille de la région métropolitaine), mais il a remarqué qu'un plus grand nombre de changements survenaient lorsque certaines données étaient imputées. L'absence de lien notable indique qu'il faut essayer de trouver d'autres façons d'expliquer ce phénomène.

Les estimations de flux bruts continuent de susciter beaucoup d'intérêt. Hubble et Juddkins (1986) ont mis au point un modèle permettant d'estimer le biais des estimations de flux bruts résultant des erreurs de réponse, modèle dont les paramètres sont estimés au moyen des taux d'erreur de réponse de la SIPP et du rapport des estimations de flux bruts à l'intérieur des cycles et d'un cycle à l'autre. Il est nécessaire de poser plusieurs hypothèses et de mettre sur pied un programme de réinterview permettant de produire des données précises sur les flux bruts pendant la période visée. Weidman (1987) propose des modèles linéaires qui tentent de représenter les liens existant entre les changements observés et les changements réels. Il est vrai que les modèles, qui utilisent uniquement les données d'enquête, sont simplifiés à l'extrême; toutefois, ils illustrent la nécessité d'obtenir plus de renseignements sur la structure des erreurs dont sont entachées les déclarations des prestations que touchent les répondants de la SIPP bénéficiant des programmes de transfert.

L'estimation composite est une technique qui combine les estimations produites pour la période en cours et celles produites pour les périodes antérieures afin d'améliorer les estimations en mettant à profit les corrélations qui existent entre les réponses fournies par les mêmes unités d'analyse à deux moments différents. Ce genre d'estimation est particulièrement efficace lorsque la corrélation est forte, ce qui est le cas pour un bon nombre d'éléments d'information importants recueillis au moyen de la SIPP. Chakrabarty (1986) a fait une première étude des types d'estimateurs composites susceptibles de convenir à la structure des données de la SIPP. Toutefois, le contenu de cette enquête n'a pas été suffisamment stable pendant les premières années de son existence pour qu'on puisse sérieusement penser à adopter un estimateur composite.

Une autre solution envisagée en vue de réduire la variance est l'utilisation de données de sources administratives pour une stratification a posteriori. Les méthodes d'estimation transversale utilisées présentent un ajustement du second degré visant à accroître la précision des estimations. À cette fin, on procède à un ajustement par quotient des estimations pour les mois de collecte et les mois de référence en fonction des estimations de population. Cependant, le Census Bureau a accès à certains fichiers de l'Internal Revenue Service et de la Social Security Administration qui peuvent être utilisés pour la production de distributions détaillées du revenu brut ajusté selon l'âge, la race et le sexe. Nous venons tout juste de commencer à étudier de quelle façon ces données administratives pourraient servir à une stratification a posteriori visant à améliorer les estimations du revenu moyen et médian des particuliers et des ménages ainsi que les estimations des déciles de la distribution du revenu des particuliers et des ménages. La question primordiale actuellement à l'étude est de savoir de combien on peut réduire de la variance des estimations avec une telle méthode. Fay et Huggins (1988) en donneront quelques indications.

Echantillonnage de sous-populations spéciales

On dit souvent que certains sous-groupes de la population sont plus touchés que les autres par la politique sociale adoptée par le gouvernement. On pense notamment aux pauvres, aux personnes âgées, aux noirs, aux citoyens d'origine latino-américaine et aux bénéficiaires des programmes fédéraux de sécurité du revenu. Dès le départ, dans l'ISDP, on voulait choisir un plan de sondage qui permettrait de rendre les estimations relatives aux sous-populations plus fiables. C'est pourquoi on a mis l'accent sur le prélèvement des échantillons dans les fichiers administratifs. On a donc souvent tiré ces derniers des listes de bénéficiaires de programmes administratifs par le gouvernement fédéral ou par les États (Kasprzyk 1983; Bowie et Kasprzyk 1987).

Un groupe de travail du Census Bureau a analysé des propositions de souséchantillonnage visant à surreprésenter certaines populations spéciales. Il s'agissait de comparer la fiabilité des estimations produites lors de divers plans de sous-échantillonnage étaient utilisés. Le groupe a mis en évidence certaines caractéristiques de sous-échantillonnage en fonction de variables démographiques et de variables relatives au revenu et il a évalué la fiabilité selon différents taux et caractéristiques de sous-échantillonnage (U.S. Bureau of the Census 1985).

Le groupe de travail en est venu à la conclusion que, pour une enquête générale sur le revenu comme la SIPP, les propositions de sous-échantillonnage ne permettaient d'améliorer que légèrement les estimations relatives aux personnes à faible revenu et que ce faible avantage ne compensait pas sur les inconvénients, notamment une plus grande complexité des opérations, le fait que l'échantillon ne soit plus autopondéré et une baisse de précision en ce qui a trait aux personnes à revenu moyen.

de la non-réponse à l'un ou l'autre des cycles du panel. Un aspect important des travaux réalisés jusqu'à présent a donc été l'étude de méthodes de compensation de la non-réponse. Kalton (1983) passe en revue les méthodes actuellement utilisées dans la recherche sur les techniques d'enquête, tandis que Kalton et Kasprzyk (1982, 1986) se penchent sur les méthodes d'imputation et résument les caractéristiques du biais et de la variance dans plusieurs classes de méthodes. Les données tirées de la SIPP peuvent être traitées soit comme des données longitudinales, soit comme des données transversales. Chapman, Bailey et Kasprzyk (1986) décrivent les méthodes utilisées pour compenser la non-réponse des unités pour la SIPP ainsi que pour d'autres enquêtes du Census Bureau. Ils parlent des problèmes que cela pose lorsqu'on a affaire à une enquête à plusieurs interviews. Dans une enquête par panel, cependant, la non-réponse peut prendre deux formes: non-réponse aux questions, qui se produit lorsqu'une unité participe à l'enquête mais ne répond pas à toutes les questions, et non-réponse aux cycles, qui se produit lorsqu'une unité participe à certaines interviews, mais pas à toutes. Heeringa et Lepkowski (1986) décrivent des classes générales de méthodes d'imputation longitudinale qui pourraient être utilisées à la place de la méthode du "hot deck", qui est une méthode d'imputation transversale. Ils comparent les résultats obtenus au moyen d'une méthode simple d'imputation longitudinale, la substitution directe longitudinale (où l'on met à la place des réponses manquantes à une interview les réponses fournies aux mêmes questions à une autre interview) aux résultats obtenus par la méthode du "hot deck". Ils montrent, et cela ne surprend personne, que la méthode de substitution directe entraîne une sous-estimation des changements. Ils concluent cependant que cette sous-estimation est peut-être préférable à la surestimation marquée qui découle de l'utilisation de la méthode du "hot deck".

Les enquêtes par panel soulèvent un problème de données manquantes que les autres types d'enquêtes ne posent pas, et c'est le problème de la non-réponse aux cycles. En général, la non-réponse aux cycles donne lieu à un plus grand nombre de données manquantes pour un individu en particulier que la non-réponse aux questions. Par ailleurs, les données obtenues auprès de répondants qui se prêtent à tous les cycles d'interviews fournissent plus de détails sur l'unité non répondante que les données obtenues auprès de répondants ayant sauté certains cycles. Pour compenser la non-réponse, il faut donc avoir recours à la pondération ou à l'imputation ou encore à une combinaison des deux. Kalton, Lepkowski et Lin (1985) parlent de ce problème et des résultats obtenus à la suite d'essais empiriques effectués dans le contexte de l'ISDP. Il ressort clairement de leur étude que le choix à faire entre la pondération et l'imputation pour régler ce genre de problème de données manquantes est loin d'être évident. Kalton (1986) de même que Kalton et Miller (1986) poussent plus loin l'analyse du problème et en viennent à la conclusion que l'imputation fausse certaines estimations et que la pondération constitue éventuellement une meilleure solution pour de grandes catégories si une diminution de la taille réelle préférable lorsqu'il s'agit d'estimations produites pour des petites catégories et que la diminution de la taille d'échantillon est considérable. Lorsqu'on travaille avec un fichier longitudinal de la SIPP contenant les données tirées de trois interviews, la différence dans la taille de l'échantillon selon qu'on procède à la pondération ou à l'imputation n'étant pas considérable, la pondération constitue la solution générale la plus sage. Enfin, Lepkowski (1988) conclut après d'autres travaux de recherche empiriques que pour régler le problème de la non-réponse aux cycles il faut absolument tenir compte de facteurs comme les principaux buts visés par le plan de sondage, le plan d'échantillonnage du panel et la répartition de la non-réponse aux cycles. Il propose des critères à retenir au moment de mettre au point une méthode de compensation et termine en disant que la pondération semble être la meilleure méthode

Réduction de l'erreur d'échantillonnage au moyen de techniques d'estimation

Deux méthodes visant à réduire l'erreur d'échantillonnage au moyen de techniques d'estimation sont à l'étude: l'utilisation d'un estimateur composite et celle de dossiers administratifs.

aux ménages selon les types auxquels ceux-ci appartiennent. Si cette constatation est confirmée par ailleurs, ce seront d'autres critères comme la facilité d'application et la simplicité des opérations qui détermineront le choix d'une définition du ménage qui tienne compte du facteur temps.

Estimation statistique d'après des définitions tenant compte du facteur temps

Les travaux de recherche sur la production d'estimations en fonction de définitions qui tiennent compte du facteur temps se sont faits dans deux directions: estimations longitudinales concernant les personnes et estimations longitudinales concernant les ménages (unité familiale ou unité de programme). Les travaux sur les estimations concernant les personnes comprennent le calcul des probabilités de sélection permettant d'obtenir des estimations longitudinales non biaisées des caractéristiques individuelles et l'utilisation de totaux de contrôle à d'autres étapes de l'estimation (Judd et coll. 1984). Kobilarcik et Singh (1986) ont raffiné ces opérations et décrit une méthode qu'ils proposent et qui permettrait de produire des poids longitudinaux pour l'analyse des données sur les personnes obtenues au cours des trois premières interviews de la SIPP.

Kobilarcik et Singh définissent l'univers longitudinal comme étant constitué de la population hors établissement institutionnel (y compris les casernes) au 1er décembre 1983, soit au milieu de la période d'interview du cycle I. L'échantillon tiré de l'univers longitudinal compte, quant à lui, les personnes admissibles qui habitent dans le logement sélectionné au moment de la première interview. Pour les besoins de cette méthode d'estimation, les personnes "interviewées" sont celles qui s'étaient prêtées aux trois premières interviews de la SIPP et qui, au moment de la première interview, faisaient partie d'un ménage dont tous les membres admissibles avaient fait l'objet d'une interview, de même que celles qui faisaient partie d'un ménage interviewé pendant le cycle I mais qui étaient décédées ou avaient déménagé à l'extérieur des limites géographiques de l'enquête dans le courant du deuxième ou du troisième cycle.

Pour la méthode d'estimation, les personnes non interviewées sont donc celles qui, au moment de la première interview, faisaient partie d'un ménage dont au moins un des membres ne s'était pas prêté à l'interview et celles qui faisaient partie d'un ménage interviewé pendant le cycle I mais qui ne s'étaient pas prêtées soit à la deuxième soit à la troisième interview, soit aux deux. On a attribué un poids positif aux personnes classées dans la catégorie des personnes interviewées. Pour cet univers, les poids ont été obtenus de la manière habituelle: en prenant l'inverse de la probabilité de sélection, en calculant un facteur de correction pour les non-interviews et en ajustant les chiffres au moyen des totaux de contrôle de la population. L'ajustement pour la non-réponse s'est faite en deux étapes, soit en ajustant d'abord pour la non-réponse des ménages et ensuite pour celle des personnes, la seconde étape faisant appel aux renseignements recueillis à la première interview.

La question des estimations longitudinales concernant les ménages (unité familiale ou unité de programme) est également à l'étude. Plusieurs façons d'aborder cette question ont été décrites par Ernst, Hubble et Judd (1984) et plus récemment par Ernst (1988). Ce dernier démontre pourquoi la pondération au moyen de l'inverse de la probabilité de sélection ne produit pas, en général, de bonnes estimations longitudinales relatives aux ménages et aux familles et il propose une classe de méthodes qui le font. Il décrit en outre les difficultés qu'on peut rencontrer si, au moment d'appliquer ces méthodes, on n'a pas accès aux données nécessaires pour établir les poids. Il définit alors les conditions que les définitions doivent satisfaire pour permettre de contourner ces problèmes. Enfin, il parle de méthodes visant à corriger les estimations longitudinales pour tenir compte de la non-réponse: ainsi qu'à ajuster les variables démographiques aux estimations indépendantes.

Non-réponse et imputation

Dans le cas d'enquêtes longitudinales comme la SIPP et celles qui ont été menées au cours de l'ISDP, les problèmes de refus et de non-réponse sélective s'ajoutent au problème cumulatif

changements susceptibles d'améliorer le champ d'observation de la SIPP à l'avenir. Kalton et Lepkowski (1985) traitent également de la règle de conduite adoptée à l'égard des répondants de la SIPP ayant déménagé et ils proposent un programme de recherche visant à mesurer le sous-recouvrement pour certaines sources et à déterminer quels sous-groupes sont les plus affectés. Plus récemment, Jean et McArthur (1987), dans une étude des données obtenues pendant cinq cycles d'interviews, ont constaté que parmi les personnes ayant déménagé après la première interview, c'est-à-dire entre le cycle 2 et le cycle 5, 69% se sont prêtées à chacune des cinq interviews, 23% n'ont pas passé la cinquième et 9% ont passé la cinquième mais en ont manqué au moins une auparavant.

4. DÉFINITIONS, PLAN DE SONDAGE ET ESTIMATION

Dans le cadre de l'ISDP et encore dans celui du programme de la SIPP, il s'est fait des travaux de recherche importants dans le domaine de la définition d'unités annuelles d'analyse à l'aide de données infra-annuelles et de l'évaluation de ces définitions d'un point de vue statistique. Le traitement de la non-réponse dans les enquêtes par panel a également suscité énormément d'intérêt parmi les chercheurs. Enfin, on a étudié et on étudie toujours les techniques d'estimation susceptibles de réduire l'erreur d'échantillonnage de même que les méthodes d'échantillonnage des sous-groupes les mieux adaptées à l'enquête.

Définitions qui tiennent compte du facteur temps

Les statistiques annuelles sur les familles et les ménages sont des indicateurs importants du bien-être économique de la nation. La SIPP permet de recueillir des données infra-annuelles, en fait des données mensuelles, qui témoignent de l'évolution de la composition des ménages. Grâce à ces données, il est possible de produire des statistiques annuelles sur les ménages qui rendent bien compte des changements que subissent ces derniers dans le courant d'une année, contrairement aux statistiques actuelles qui ne font tout simplement pas cas de cette réalité. La construction d'unités annuelles d'analyse, qu'ils s'agisse de ménages, de familles ou d'unités de programme, pose le problème des poids longitudinaux et des techniques d'imputation à adopter. Le principal problème, cependant, est conceptuel. Etant donné que la composition des ménages varie entre le début et la fin d'une année, quand est-il indiqué pour les mesures annuelles de rendre compte de cette évolution? Autrement dit, comment faudrait-il définir les ménages et les familles de manière à pouvoir tenir compte des données recueillies à deux moments ou plus et à ne pas être en contradiction avec les estimations transversales habituellement produites sur les ménages et les familles?

Les analystes du Census Bureau se sont longuement penchés sur la question d'une définition des ménages et des familles qui tiendrait compte du facteur temps (McMillen et Herriot 1985; Citro 1985). Citro, Hernandez et Herriot (1986) ont fait état des travaux de recherche empirique dans lesquels plusieurs définitions de ménages de ce genre et diverses mesures du revenu annuel et des types de famille ont été examinées. Dans ces travaux, quatre définitions possibles ont été mises de l'avant: on considère qu'un ménage est le même au bout d'une certaine période (1) si la personne de référence est la même; (2) si le membre principal du ménage est le même (la différence entre cette définition et la précédente se manifeste dans le traitement des ménages comportant un couple marié, où la personne de référence peut être soit la femme soit le mari mais où le membre principal est toujours la femme); (3) si la personne de référence est la même et si la famille qui le constitue appartient au même type; (4) si la personne de référence est la même, la famille qui le constitue appartient au même type et le nombre de membres n'a pas changé.

Il ressort à première vue de ces travaux que le choix d'une définition n'influe pas sensiblement sur les statistiques annuelles relatives aux ménages à faible revenu ni sur celles relatives

Olson (1980) décrit certains résultats de la première expérience. Fait non surprenant, lorsque la période visée par les questions est de six mois, la proportion du revenu déclarée pour le début de la période est inférieure à ce qu'elle devrait être, et ce pour plusieurs sources de revenu comme la rémunération, l'aide aux familles à faible revenu avec enfants à charge et les prestations d'assurance-chômage. Cette observation, bien que non définitive, confirme l'hypothèse selon laquelle les chances d'omission par défaut de mémoire augmentent avec la durée de la période de référence. Il est ressorti d'une autre analyse que le nombre de sources de revenu déclarées par ménage pour les trois premiers mois de la période de référence de six mois était inférieur au nombre obtenu pour les mêmes mois lorsque la période de référence était de trois mois. Les résultats de la deuxième expérience n'ont pas été analysés à cause du retrait des subventions accordées au programme de mise au point de l'enquête. Les résultats de la première expérience et certains faits observés à d'autres étapes de l'ISDP ont amené les planificateurs de la SIPP à adopter une période de référence de quatre mois pour cette enquête; cette décision permet de maintenir les coûts dans les limites permises par le budget et la qualité des données à un niveau satisfaisant.

Règle de conduite adoptée à l'égard des répondants ayant déménagé

Un aspect important du plan de sondage utilisé dans l'ISDP et à présent dans la SIPP est le fait que toutes les personnes faisant partie d'un ménage de l'échantillon au moment de la première interview continuent à faire partie de l'échantillon pendant les deux années et demie que le panel est conservé, même si une personne ou plus déménage. Pour des raisons d'ordre financier et opérationnel, il a été décidé que les interviews sur place seraient menées à la nouvelle adresse uniquement si certaines contraintes géographiques étaient respectées; dans l'ISDP, il fallait que la nouvelle adresse se situe dans un rayon de 50 miles d'une unité primaire d'échantillonnage et, dans la SIPP, dans un rayon de 100 miles. Lorsqu'on crée un panel, on tire un échantillon d'adresses, et ce n'est qu'à la première interview que les personnes habitant à chaque adresse sont identifiées. Après cela, l'échantillon n'est plus un échantillon d'adresses mais un échantillon de personnes, et il est constitué de tous les individus énumérés lors de la première interview. On demande à ces derniers et à quiconque partage le même logement de se prêter aux interviews ultérieures.

Pendant la durée de l'ISDP, deux questions relatives aux répondants ayant déménagé ont été considérées importantes: 1) la production d'estimations transversales à chaque interview et 2) les coûts supportés lorsqu'on suit ces répondants. Huang (1984) propose plusieurs poids de base non biaisés pour les estimations transversales relatives à la population hors établissement institutionnel à utiliser si l'échantillon contient des répondants ayant déménagé. Il établit le lien entre les observations effectuées à un moment donné et les probabilités de sélection des ménages de l'échantillon original. Deux approches sont décrites dans son étude: la première, fondée sur la multiplicité, est liée au nombre de façons dont un nouveau ménage peut se retrouver dans l'échantillon et la seconde, fondée sur la notion de "part égale", suppose que tous les membres d'un ménage ont le même poids. Pour la SIPP, comme pour l'ISDP, on a adopté l'approche fondée sur la notion de "part égale".

La question des coûts a été traitée dans une étude consacrée à ce sujet ("Mover's Cost Study"). Cette étude avait pour but d'évaluer les coûts, au chapitre de la collecte des données, reliés au fait de suivre les répondants après qu'ils ont déménagé. White et Huang (1982) font un compte rendu de l'étude et présentent certains résultats fondés sur la ligne de conduite adoptée, au cours du test pilote, à l'égard de ces répondants. Ils ont trouvé que lorsqu'on a suivi ces derniers pendant un an, le nombre de ménages devant être interviewés a augmenté de 8.8%; en outre, au bout de quinze mois, les répondants ayant déménagé représentaient 22% de l'échantillon total, le nombre d'interviews s'était accru de 7% et le nombre de miles déclarés par les intervieweurs, de 11.4%.

Jean et McArthur (1984) se sont penchés sur la question de la collecte des données relatives aux répondants ayant déménagé dans le contexte de la SIPP et ils recommandent certains

cadre de l'ISDP, on a tenté de déterminer si les renseignements relatifs à des étudiants absents de chez leurs parents obtenus auprès de répondants substitués étaient exacts. Dans un premier temps, on a mené l'interview par personne interrogée auprès d'un membre du ménage des parents et, dans un deuxième temps, on a mené l'interview auprès de l'étudiant lui-même à l'endroit où il demeurait pendant la période de ses études. Les résultats de cette étude sont décrits par Roman et O'Brien (1984). L'analyse est limitée à cause d'erreurs commises sur le plan de l'organisation et de la réalisation du test. Les auteurs signalaient cependant que les répondants substitués ne sont souvent pas en mesure de dire si un étudiant touche une certaine catégorie de revenu et que, même s'ils sont en mesure de le faire, lorsqu'on leur demande plus de détails, ils sont portés à répondre qu'ils ne les connaissent pas. Les auteurs notent également que plus le montant du revenu ou de la dépense est élevé, plus la réponse fournie par un répondant substitut est bonne.

Mode de collecte des données

La plupart des interviews (environ 95%) menées pour la SIPP sont des interviews sur place (Kalon, McMillen et Kaspzyk 1986). Comme ce genre d'interview coûte de plus en plus cher, le Censur Bureau étudie actuellement la possibilité d'augmenter substantiellement le nombre d'interviews téléphoniques réalisées pour cette enquête. On a donc effectué au mois de juin 1985 un essai préliminaire visant à déterminer la faisabilité d'interviews téléphoniques 'préparées', c'est-à-dire d'interviews téléphoniques menées auprès de ménages qui avaient accordé des interviews sur place lors d'un cycle antérieur. Cet essai préliminaire a été réalisé par deux bureaux régionaux du Censur Bureau auprès d'un échantillon de 280 ménages. Le taux de refus (environ 2,5%) et le taux de non-contact (environ 11%) étaient conformes aux attentes du personnel. En ce qui concerne les taux de non-réponse aux diverses questions, ils n'étaient pas plus élevés que prévu non plus (U.S. Bureau of the Census, 1986). Étant donné les résultats obtenus, on a procédé à un essai national de l'interview téléphonique pour la SIPP (SIPP National Telephone Test) qui s'est déroulé du mois d'août au mois de novembre 1986 et du mois de février au mois d'avril 1987. Il s'agissait d'évaluer l'utilisation sur une grande échelle des interviews téléphoniques 'préparées' et de déterminer si les gens sont prêts à fournir des données par téléphone au cours de deux interviews consécutives. On a décidé que les ménages situés dans 50% des segments feraient l'objet d'un nombre maximum d'interviews téléphoniques et les autres, d'un nombre maximum d'interviews sur place. Les interviews ont effectué presque toutes les interviews téléphoniques à domicile. Gbur et Durant (1987) donnent les résultats préliminaires de la première étape de l'expérience. Selon eux, l'interview téléphonique ne semble pas avoir eu de conséquences graves sur les taux de réponse des ménages, et les taux de réponse des personnes étaient semblables quel que soit le mode utilisé. Quant aux taux de non-réponse aux questions, ils n'ont été que légèrement influencés par le recours à l'interview téléphonique. Le compte rendu des autres résultats suivra.

Durée de la période de référence

L'ISDP ayant porté principalement sur les techniques de collecte des données susceptibles d'améliorer la déclaration des revenus monétaires et non monétaires, la durée de la période de référence pour la plupart des questions posées au cours de l'enquête est un sujet auquel on a accordé beaucoup d'importance lors de la détermination du plan de sondage. On a abordé ce sujet de deux points de vue pendant le programme. Premièrement, on a comparé les résultats obtenus au moyen d'une interview dont la période de référence était de six mois et ceux obtenus au moyen de deux interviews consécutives dont la période de référence était de trois mois chacune. Deuxièmement, on a comparé les résultats obtenus si l'on demandait aux répondants de déclarer les revenus de biens qu'ils avaient touchés au cours des trois mois antérieurs et ceux obtenus si on leur demandait de déclarer les revenus touchés au cours des six mois antérieurs.

3. COLLECTE DES DONNÉES

Quatre aspects de la collecte des données pour la SIP seront traités ici: 1) les règles de conduite adoptées à l'égard des répondants; 2) le mode de collecte des données; 3) la durée de la période de référence; 4) les règles de conduite adoptées à l'égard des répondants ayant déménagé.

Règles de conduite adoptées à l'égard des répondants

Lorsqu'on a à mener des entretiens dans des ménages composés de plusieurs membres, il faut décider si les réponses par personne interposée sont acceptables. Comme il peut arriver que les membres du ménage ne soient pas tous présents au moment de l'interview, on peut gagner du temps et de l'argent en posant aux personnes qui sont là les questions concernant celles qui n'y sont pas. Toutefois, pour certaines questions, les données fournies par un répondant substitut risquent d'être moins précises que celles qui auraient été communiquées par le répondant lui-même. Kalton, Kasprzyk et McMillen (1988) traitent de cette question dans le contexte des enquêtes par panel.

Une expérience portant sur les règles de conduites adoptées à l'égard des répondants a été menée dans le cadre de l'ISDP. On a comparé la qualité des données obtenues dans un groupe témoin où une déclaration par personne interposée a été acceptée si un des membres du ménage se sentait en mesure de fournir les renseignements concernant la personne absente, aux données obtenues dans un autre groupe où les déclarations par personne interposée n'ont été acceptées que dans des situations exceptionnelles (par exemple si le répondant était physiquement ou mentalement incapable de répondre aux questions, s'il ne parlait pas l'anglais ou s'il allait être absent pendant toute la durée de l'enquête). Environ 85% des adultes interrogés dans les ménages où l'on a exigé l'autodéclaration parlaient en leur propre nom comparativement à 65% dans les ménages où l'on a accepté des déclarations par personne interposée. On a donc obtenu environ 20% de déclarations en nom propre de plus dans le groupe où ce genre de déclaration a été exigé que dans l'autre groupe (Coder 1980).

Les taux de refus étaient légèrement plus élevés dans le groupe où l'on a exigé l'autodéclaration tandis que le pourcentage de ménages interviewés était un peu plus fort dans le groupe où l'on a accepté les déclarations par personne interposée. L'écart était cependant trop petit dans un cas comme dans l'autre pour qu'on puisse en déduire laquelle des deux règles de conduite était la plus avantageuse. La proportion de personnes non-interviewées dans les ménages où au moins un adulte avait répondu aux questions était plus élevée dans le groupe où l'on a exigé l'autodéclaration. Pour ce qui est des pourcentages de personnes ayant touché diverses catégories de revenu, la différence observée entre les deux groupes était légère également, et les taux de non-réponse due à l'impossibilité ou au refus de communiquer les renseignements demandés aux questions sur les montants touchés dans chaque catégorie de revenu n'étaient pas systématiquement plus faibles dans le groupe où l'autodéclaration a été exigée.

Dans le groupe où l'on a exigé l'autodéclaration, les personnes interrogées ont plus souvent consulté leurs dossiers avant de répondre aux questions concernant la rémunération, et les taux de réponse à la question sur la rémunération horaire étaient plus élevés, mais en général l'écart entre les résultats obtenus dans les deux groupes n'était pas concluant. C'est pour cette raison et parce qu'on a estimé qu'il coûtait de 4 à 6% plus cher d'exiger l'autodéclaration de même que parce qu'il a fallu adopter une procédure de rappel pour obtenir des renseignements essentiels qui n'étaient pas connus au moment de l'interview que les déclarations par personne interposée sont maintenant acceptées dans la SIP.

Il a également fallu se demander quelle règle de conduite adopter à l'égard des étudiants universitaires. En général, on considère qu'un étudiant fait partie du ménage de ses parents tant qu'il ne s'est pas établi ailleurs de façon permanente. On traite donc habituellement les étudiants qui n'habitent pas chez leurs parents pendant la durée de leurs études comme des membres du ménage temporairement absents, et on demande à un autre membre du ménage de leurs parents de répondre aux questions à leur place. Lors d'un test pilote effectué dans le

à un an d'intervalle. Plus particulièrement, les données sur l'actif et le passif, recueillies au cours du cycle 4 après du panel de 1984, ont été fournies à la moitié des répondants lors de l'interview du cycle 7. Ces données n'ont été communiquées qu'à la moitié des répondants afin de pouvoir comparer l'approche avec rétro-information à l'approche indépendante.

L'argument invoqué pour justifier l'approche avec rétro-information est que les répondants donnent de meilleures estimations de l'évolution de leur revenu si on leur rappelle auparavant le montant qu'ils avaient déclaré l'année précédente. Si les répondants savent de combien leur actif a changé et si on leur rappelle le montant de départ, on peut supposer que le montant déclaré pour l'année en cours reflète le changement réel pendant la période de référence. Lamas et McNeil (1987) analysent ces données, mais ils n'affirment rien de définitif sur l'incidence de l'approche avec rétro-information car ils ne disposent pas de données de référence. Ils disent bien, cependant, que l'interview avec rétro-information ne change pas les estimations transversales et que les différences observées entre les sous-groupes pour ce qui est de la valeur nette sont celles auxquelles on s'attendait. L'étude de l'évolution de la valeur nette au microniveau dans les ménages ayant fourni toutes les données sur leurs ressources tend à montrer que l'interview avec rétro-information fait baisser les estimations des changements de la valeur nette.

La question des résultats obtenus selon qu'on procédait à une interview indépendante ou à une interview avec rétro-information s'est également posée relativement à la mesure de la branche d'activité et de la profession. Après des panels de 1984 et de 1985, ces données ont été recueillies indépendamment à chaque interview même si le répondant n'avait pas changé d'employeur. Cette façon de procéder tient compte du fait que les fonctions d'un travailleur peuvent varier à l'occasion et permet de noter cette évolution. En outre, les fonctions peuvent changer au point que le poste occupé par le travailleur en question n'est plus classé dans la même catégorie d'une interview à l'autre même si celui-ci est au service du même employeur. La collecte indépendante des données sur la branche d'activité et la profession pose cependant des problèmes. Une légère variation dans la manière dont le répondant décrit ses fonctions ou dans l'interprétation de cette description par les personnes chargées d'attribuer les codes de classification peut entraîner un changement de catégorie professionnelle injustifié. L'étude de ce problème a permis de produire des estimations du nombre de fois que la profession et la branche d'activité des répondants qui étaient restés au service du même employeur avaient été classifiées dans une catégorie différente d'une interview à l'autre. Chez les personnes faisant partie du panel de 1984 qui avaient déclaré travailler pour le même employeur au cours des douze premiers mois de l'enquête, on a observé un changement du code à trois chiffres d'une interview à l'autre dans environ 40% des cas pour ce qui est de la profession et dans 20% des cas pour ce qui est de la branche d'activité (Kaltou, McMillen et Kasprzyk 1986).

Pour réduire le nombre de changements de codes de profession et de branche d'activité découlant d'une erreur de réponse aléatoire ou d'une mauvaise interprétation des réponses par les codeurs et pour abréger les interviews, on a décidé de modifier la SIPP à partir du panel de 1986. En l'occurrence, on a introduit une question de sélection par laquelle on demande aux répondants si leurs activités ou fonctions ont changé au cours des huit derniers mois. Si la réponse est négative, les questions détaillées sur la profession et la branche d'activité ne leur sont pas posées. Les codes attribuables à ces deux variables sont alors assignés à partir des réponses fournies à l'interview précédente. Il importe de signaler le fait suivant: bien que ce changement ait été introduit au moment où le panel de 1986 a été créé, les données sur la profession et la branche d'activité se rapportant au panel de 1985 qui ont été recueillies au cours de la même période que celles se rapportant au panel de 1986 l'ont été selon l'interview indépendante. Ces deux façons de procéder utilisées simultanément constituent donc une expérience en quelque sorte spontanée dont les résultats n'ont pas encore été analysés.

cherchait pas à obtenir le montant reçu relativement à une catégorie de revenu tant que toutes les catégories pour lesquelles un revenu avait été touché n'avaient pas été déterminées.

L'hypothèse que l'on cherchait à vérifier était que les renseignements sur le revenu fournis au moyen du questionnaire long seraient plus complets et plus précis; Olsson (1980) résume l'analyse qui a été faite sur les deux versions du questionnaire. Cette analyse a pris plusieurs formes, dont il est question dans le résumé: 1) observation des intervieweurs pendant la formation et les interviews; 2) compte rendu des intervieweurs et des observateurs; 3) examen de chaque questionnaire complet; 4) analyse des taux de participation à l'enquête et des taux de réponse aux questions; et enfin 5) analyse de la qualité des données et des questionnaires rejetés à la vérification, particulièrement ceux dont le rejet était dû au fait que l'interviewer n'avait pas suivi les instructions indiquant de "passer à". C'est une version modifiée du questionnaire long qui a été retenue pour la suite des essais et finalement pour la SIP. Les intervieweurs et les répondants ont en effet trouvé que le questionnaire long était plus facile à remplir et, grâce à lui, on a obtenu des taux de déclaration du revenu plus élevés.

Toujours dans le cadre de l'ISDP, on a fait une expérience mettant en jeu différents types de questionnaires. Cette expérience opposait deux approches: l'une axée sur le ménage et l'autre axée sur l'individu, cette dernière faisant appel à un questionnaire dont la forme et le contenu étaient le fruit de tests pilotes réalisés antérieurement. Dans l'approche axée sur le ménage, on s'est servi d'une version remaniée d'un questionnaire utilisé pour l'essai du supplément sur le revenu de la CPS réalisé en avril 1978. On cherchait, au moyen de ce questionnaire, à réduire le fardeau de réponse en demandant à un seul membre du ménage s'il y avait dans le ménage quelqu'un qui avait touché une certaine catégorie de revenu au cours de la période de référence. Chaque réponse affirmative était suivie d'une question pour déterminer quel membre du ménage avait touché le revenu en question. Ce n'est qu'une fois qu'on avait obtenu des renseignements au sujet de toutes les catégories de revenu perçu par l'ensemble des membres du ménage qu'on cherchait à connaître le montant reçu dans chaque cas. On pensait que cette approche aurait pour effet de réduire la durée de l'interview sans faire baisser la qualité des données.

Dans l'approche axée sur l'individu, on posait des questions sur toutes les sources de revenu à un premier membre du ménage, puis à un second, et ainsi de suite. On remplissait un questionnaire distinct pour chaque adulte faisant partie du ménage échantilloné, mais on avait recours autant que possible à des instructions indiquant de "passer à" et à des cases qu'il suffisait de cocher afin de diminuer le nombre de questions posées à chaque répondant.

Du point de vue de la qualité des données et de la durée de l'interview, les différences observées entre les deux approches se sont avérées minimes. Elles n'étaient pas grandes non plus en ce qui a trait aux estimations du pourcentage de personnes ayant touché diverses catégories de revenu ni du nombre de cas de non-réponse dus à l'impossibilité ou au refus de fournir les renseignements demandés. On s'attendait à ce que la durée de l'interview soit beaucoup moins grande lorsque l'approche axée sur le ménage était utilisée plutôt que l'approche axée sur l'individu, or elle n'a été inférieure que d'environ cinq minutes par ménage et d'environ trois minutes par personne. Comme l'approche axée sur le ménage n'a pas donné de résultats sensiblement meilleurs que l'approche axée sur l'individu, c'est cette dernière qui, légèrement améliorée et raffinée, a été retenue pour la SIP.

Le type de questionnaire à utiliser et la marche à suivre pour la collecte des données font l'objet de discussions qui se poursuivent dans le cadre du programme de la SIP. La principale question consiste à savoir si l'on obtient de meilleures estimations au moyen d'interviews ne faisant pas appel aux réponses données antérieurement (interviews dites indépendantes) ou au moyen d'interviews au cours desquelles on rappelle aux répondants les renseignements qu'ils ont déjà fournis (interviews dites avec rétro-information). Dans la SIP, on a recours à l'approche avec rétro-information pour mettre à jour les données à chaque interview et ainsi connaître les dernières tendances relatives au revenu, mais cette approche n'a pas été évaluée. Il est possible d'adopter l'approche avec rétro-information pour la collecte des données sur la valeur nette personnelle. Dans la SIP, ces données sont des données ponctuelles recueillies

l'instrument et aux méthodes d'enquête, il était difficile d'y remédier. En 1975, le Département américain de la santé et des services sociaux a donc décidé de mettre sur pied l'ISDP pour corriger les principaux défauts de la CPS, en l'occurrence: 1) la sous-déclaration du revenu de biens et d'autres sources moins fréquentes de revenu, 2) les erreurs de déclaration et de classification ayant trait aux prestations des principaux programmes de sécurité du revenu et aux autres types de renseignements que les gens ont du mal à fournir correctement (par exemple la répartition mensuelle du revenu touché au cours de l'année) et 3) le manque d'informations nécessaires à l'analyse de la participation et de l'admissibilité aux programmes. Les tests pilotes effectués dans le cadre de l'ISDP se distinguent des autres enquêtes, particulièrement la CPS, à plusieurs égards, notamment: 1) les interviews ont été menées à intervalles réguliers dans le courant de l'année; 2) la plupart des catégories de revenu ont été déclarées sur une base mensuelle; 3) le revenu a été déclaré sur une base individuelle; 4) on a suivi les individus tout au long de la période d'enquête afin d'obtenir des données sur l'évolution des revenus et de la composition des familles; 5) on a recueilli des données sur des sujets particuliers comme l'incapacité, la garde des enfants, la fécondité, la valeur nette et les impôts payés afin d'en savoir davantage sur le contexte dans lequel les prestations sont versées ainsi que sur les personnes qui dépendent des programmes pour vivre et le bien-être économique de la population en général. Comme l'ISDP a donné naissance à la SIPP, on retrouve dans cette dernière de nombreuses caractéristiques du premier, dont le plan de sondage ainsi que le contenu et la présentation du questionnaire.

La SIPP a été lancée en octobre 1983 à titre d'enquête permanente comportant un panel de 21,000 ménages choisis pour représenter la population hors établissement institutionnel des États-Unis. Chaque ménage est interviewé une fois tous les quatre mois pendant environ deux ans et demi, la période de référence des principales questions de l'enquête étant les quatre mois qui précèdent l'interview. Chaque ménage est donc interviewé huit fois en tout. Un nouveau panel est introduit chaque année. Ce plan de sondage permet de produire des estimations transversales à partir des réponses de deux panels. Pour plus de renseignements à propos du plan de sondage, des opérations et du contenu relatifs à la SIPP, voir Nelson, McMillen et Kasprzyk (1985).

Le présent article traite de certains aspects du programme (aspects méthodologiques, plan de sondage, aspects statistiques) qui sont des sujets de préoccupation. Nous parlerons 1) de la conception du questionnaire; 2) de la collecte des données, notamment des règles de conduite adoptées à l'égard des répondants, du mode de collecte des données, de la durée de la période de référence et des règles de conduite adoptées à l'égard des personnes ayant déménagé; 3) des définitions, du plan de sondage et des estimations; 4) de l'erreur de réponse.

2. CONCEPTION DU QUESTIONNAIRE

L'ISDP visait principalement à résoudre des problèmes reliés à la sous-déclaration et à la mauvaise classification du revenu dans le supplément de mars de la CPS. Au cours d'un test pilote de l'ISDP, on a eu recours à deux versions du questionnaire que nous désignerons, pour simplifier, par les expressions "questionnaire abrégé" et "questionnaire long". Avec le questionnaire abrégé, on cherchait à recueillir directement des données sur le revenu tout en limitant le fardeau de réponse. Il s'agissait d'interroger chaque membre du ménage sur certaines catégories de revenu qu'il aurait pu percevoir. Si, pour une catégorie donnée, il avait reçu de l'argent, le montant touché au cours de la période de référence était déterminé avant de passer à la catégorie suivante.

Par contre, avec le questionnaire long, la stratégie adoptée consistait à cerner les événements, les expériences et autres détails associés à certaines catégories de revenu perçu. Le formulaire contenait un ensemble assez complet de questions d'approfondissement concernant le revenu touché ainsi que de longues questions visant à déterminer les montants en cause. On ne

Domaines de recherche relatifs à la SIPP (enquête sur le revenu et la participation aux programmes)¹

DANIEL KASPRZYK²

RÉSUMÉ

La SIPP (Survey of Income and Program Participation/enquête sur le revenu et la participation aux programmes) est une enquête permanente menée par le U.S. Bureau of the Census auprès d'un échantillon de ménages représentatif de la population à l'échelle nationale. Le but principal de la SIPP est d'améliorer la mesure de l'information sur la situation économique des ménages et des particuliers aux États-Unis. Elles sont basées sur un questionnaire contenant des questions d'approfondissement dont la période de référence est courte. Le plan de sondage à plusieurs interviews de la SIPP soulève des questions d'ordre méthodologique et statistique qui concernent toutes les enquêtes par panel menées auprès de familles et de particuliers. Ces questions sont traitées dans le présent article du point de vue de la SIPP. Il s'agit: 1) de la conception du questionnaire; 2) de la collecte des données, notamment des règles de conduite adoptées à l'égard des répondants, du mode de collecte des données, de la durée de la période de référence et des règles de conduite adoptées à l'égard des personnes ayant déménagé; 3) des définitions, du plan de sondage et des estimations; 4) de l'erreur de réponse.

MOTS CLÉS: Enquêtes par panel; conception du questionnaire; plan de sondage; estimations longitudinales; erreur de réponse.

1. INTRODUCTION

La SIPP (Survey of Income and Program Participation/enquête sur le revenu et la participation aux programmes) est une nouvelle enquête permanente menée par le U.S. Bureau of the Census auprès d'un échantillon de ménages représentatif de la population à l'échelle nationale. Elle permet de recueillir une information complète et détaillée au sujet des ressources économiques des ménages américains et de l'incidence que les programmes fiscaux et les programmes de transfert de l'administration publique peuvent avoir sur la situation financière de ces derniers. Les responsables des politiques fédérales se basent sur les données tirées de l'enquête pour déterminer l'efficacité des programmes fiscaux et des programmes de transfert, estimer les coûts futurs des programmes et le nombre de bénéficiaires éventuels ainsi que pour évaluer les répercussions de toute modification envisagée. Le but de la SIPP est d'améliorer la mesure de l'information sur la situation économique des ménages et des particuliers aux États-Unis. C'est l'aboutissement d'un vaste programme d'élaboration, l'ISDP (Income Survey Development Program/programme de mise au point de l'enquête sur le revenu) dans le cadre duquel on s'est penché sur les définitions, les méthodes, les questionnaires, les périodes de référence et d'autres éléments de nature semblable. (Ycas et Liminger 1981).

La SIPP a été créée parce qu'il fallait combler les lacunes du supplément sur le revenu greffé chaque année au mois de mars à la CPS (Current Population Survey/enquête sur la population), lequel était jusqu'alors la principale source de renseignements sur la répartition du revenu des ménages et des particuliers aux États-Unis. Ces lacunes étant liées au plan de sondage, à

¹ Cet article contient les résultats des travaux de recherche effectués par le personnel du Bureau of the Census. Les opinions qui y sont exprimées sont entièrement celles de l'auteur et ne reflètent pas nécessairement la position du Bureau of the Census.

² Daniel Kasprzyk, SIPP Research and Coordination Staff, United States Bureau of the Census, Washington D.C. 20233

de façon substantielle le taille de l'échantillon. L'application de la méthode optimum entraîne une réduction additionnelle de l'ordre de 20% pour les trois strates figurant dans ces deux tableaux. Pour un coefficient de variation donné, la variation de la puissance " p " influe peu sur la taille de l'échantillon. Comme prévu, la taille d'échantillon augmente lorsque le coefficient de variation (c) diminue (pour une même valeur de p). La méthode optimale définit moins d'unités à tirage complet (strate 3) que la méthode composée ou, en d'autres termes, la borne de la strate à tirage complet est plus élevée pour la première méthode que pour la seconde. La règle de la racine de f cumulative n'est plus aussi efficace lorsqu'il s'agit de déterminer la borne de la strate à tirage complet. Il est facile de voir que la borne calculée à l'aide de cette règle est considérablement plus élevée que celles calculées à l'aide des autres méthodes.

Dans le tableau 3, nous comparons les résultats obtenus par la méthode composée et la méthode optimale pour deux populations lorsque le nombre de strates varie, étant donné un coefficient de variation et une valeur de p (pour la répartition proportionnelle à X à la puissance p). Des conclusions semblables à celles tirées des deux premiers tableaux se dégagent du tableau 3. L'accroissement du nombre de strates a pour effet de réduire le nombre d'unités échantillonnées, peu importe la méthode utilisée. Toutefois, la réduction est plus prononcée dans le cas de la méthode optimale.

5. CONCLUSION

Le découpage optimal d'une population asymétrique en une strate à tirage complet et en un certain nombre de strates à tirage partiel a permis de réduire sensiblement la taille de l'échantillon global pour un degré de précision relative donné. La méthode proposée peut être adaptée à n'importe quel mode de répartition et à n'importe quel nombre de strates. On peut également faire abstraction de la condition relative au tirage complet.

L'algorithme, qui est récursif par définition, converge rapidement. Il peut être appliqué facilement par ordinateur par l'intermédiaire du SAS, du FORTRAN ou de n'importe quel autre langage évolué.

BIBLIOGRAPHIE

- BANKIER, M.D. (1988). Power allocations, determining sample sizes for sub-national areas. À paraître dans *The American Statistician*.
- CARROL, J. (1970). Allocation of a sample between States. Note de service non publiée de l'Australian Bureau of Census and Statistics.
- COCHRAN, W.G. (1977). *Sampling Techniques*, (3^e éd.). New York: John Wiley & Sons.
- DALENIUS, T. (1950). The problem of optimum stratification. *Skandinaviskskrift*, 33, 203-213.
- DALENIUS, T., et GURNÉY, M. (1951). The problem of optimum stratification. II, *Skandinaviskskrift*, 34, 133-148.
- DALENIUS, T., et HODGES, J.L.Jr. (1959). Minimum variance stratification, *Skandinaviskskrift*, 54, 88-101.
- FELLEGLI, I.P. (1981). Should the census counts be adjusted for allocation purposes? — Equity considerations. Dans *Current Topics in Survey Sampling* (éds. D. Krewski, R. Platek et J.N.K. Rao), New York: Academic Press, 47-76.
- GLASSER, G.J. (1962). On the complete coverage of large units in a statistical study. *Revue Internationale de Statistique*, 30, 28-32.
- HIDIROGLOU, M.A. (1986). The construction of a self-representing stratum of large units in survey design, *The American Statistician*, 40, 27-31.
- SETHI, V.K. (1963). A note on optimum stratification of populations for estimating the population means. *Australian Journal of Statistics*, 5, 20-33.

Tableau 3

Effet de l'accroissement du nombre de strates sur la taille d'échantillon pour deux méthodes de stratification
 $p = 1, c = 0.05$

Population 1 (N = 1221) Méthode de stratification	Nombre de strates				
	3	4	5		
Strates	N_h	n_h	$b(h)$	N_h	n_h

Composée	1	1017	16	897	6
	2	152	11	465,180	194
	3	52	52	1,131,961	78
	4			52	52
	5				
	Total	79		67	

Optimale	1	858	8	704	3
	2	323	16	271,920	373
	3	40	40	1,867,254	112
	4			32	32
	5				
	Total	64		48	

Composée	1	106	6	84	2
	2	39	6	38	2
	3	16	16	23	2
	4			16	16
	5				
	Total	28		22	

Optimale	1	86	4	55	1
	2	65	9	61	3
	3	10	10	39	5
	4			6	6
	5				
	Total	23		15	

ou

$$n(t') = t' + \frac{(N - t')^2 S_2^2}{(N - t')^2 S_2^2 + (NcY)^2 + (N - t')^2 S_2^2} \quad (4.5)$$

Les tableaux 1 et 2 donnent les résultats pertinents pour une grande population (population 1) et une petite population (population 2) étant donné divers coefficients de variation et diverses valeurs de p (pour la répartition à la puissance p). Le tableau 3 donne les résultats obtenus pour la grande population (population 1) et une population de taille moyenne (population 3) lorsqu'on modifie le nombre de strates. Pour les trois tableaux, la répartition de l'échantillon entre les strates à tirage partiel s'est faite suivant le mode de répartition proportionnelle à Y à la puissance p . Les tableaux 1 et 2 nous permettent de tirer les conclusions suivantes. La règle de la racine de f cumulative est très inefficace dans les présentes circonstances. La méthode composée réduit

moyen de la formule d'approximation d'Hidiroglou (1986), puis à appliquer la règle de la racine de f cumulative pour découper la population qui n'est pas incluse dans la strate à tirage complet en un certain nombre de strates à tirage partiel. Nous désignerons respectivement ces méthodes comme la règle $f^{1/2}$ CUM et la méthode combinée; quant à l'algorithme proposé, nous le désignerons comme la méthode optimale. Il n'est pas réaliste d'utiliser uniquement la méthode de Dalenius-Hodges (1959) car, dans la pratique, celle-ci ne serait appliquée qu'une fois que la strate à tirage complet aurait été définie à l'aide d'une règle arbitraire donnée. Néanmoins, nous illustrons ici l'application de cette méthode pour mettre en garde le lecteur contre une utilisation inconsidérée de cette méthode pour des populations fortement asymétriques.

Hidiroglou (1986) détermine l'emplacement de la coupure entre les strates au moyen de la formule itérative suivante:

$$b_{TA}'' = \mu_{[N-t']} = \left\{ \frac{(N-t')^2}{N^2} c^2 \bar{y}_2 + S^2_{[N-t']} \right\}^{1/2}, \quad (4.1)$$

$$\mu_{[N-t']} = \frac{1}{N-t'} \sum_{i=1}^{N-t'} y_{(i)} \quad (4.2)$$

Tableau 1

Effet de la variation du coefficient de variation et de la puissance p sur la taille d'échantillon pour trois méthodes de stratification (population I — taille = 1221)

Méthode de stratification			
Règle $f^{1/2}$ Cum	Composée	Optimale	
N_h	N_h	N_h	$b_{(h)}$
n_h	n_h	n_h	$b_{(h)}$

0.05	0.25	1	1196	20	3,715,320	1017	16	465,180	290	891	11	302,912	1,835,930
Total		3	5	5	14,786,280	52	52	1,131,961	40	40	64	1,835,930	1,835,930
0.05	0.50	1	1196	20	3,715,320	1017	16	465,180	863	10	14	289,422	1,832,038
Total		3	5	5	17,786,280	52	52	1,131,961	40	40	64	1,832,038	1,832,038
0.01	1.00	1	1196	20	3,715,320	751	37	196,840	687	36	78	162,068	564,076
Total		3	5	5	14,786,280	255	34	383,033	374	160	160	564,076	564,076
0.05	1.00	1	1196	20	3,715,320	1017	16	465,180	858	8	16	271,920	1,867,254
Total		3	5	5	14,786,280	52	52	1,131,961	40	40	64	1,867,254	1,867,254
0.10	1.00	1	1196	20	3,715,320	1073	7	592,900	1007	7	9	442,357	4,032,950
Total		3	5	5	14,786,280	39	4	1,953,113	191	23	23	4,032,950	4,032,950

*Une répartition disproportionnée est nécessaire pour satisfaire au coefficient de variation.

$$\bar{Y}_h = \frac{1}{N_h} \sum_{j=b(h)}^{f=b(h-1)+1} Y_{(j)} \tag{3.13}$$
$$S_h^2 = \frac{1}{N_h} \sum_{j=b(h)}^{f=b(h-1)+1} Y_{(j)}^2 - N_h \bar{Y}_h^2 \tag{3.14}$$

pour $h = 1, \dots, L$.

Ces formules peuvent alors nous permettre de déterminer sans difficulté à l'aide de la méthode

itérative suivante les bornes de strates $b^{(1)}, \dots, b^{(L-1)}$ de manière que la taille n de l'échantillon

global minimum pour un degré de fiabilité (c) et un mode de répartition données:

ÉTAPE 0 : Classer les éléments de la population Y_1, \dots, Y_N par ordre croissant et poser $b^{(0)} = Y^{(1)}$ et $b^{(L)} = Y^{(N)}$.

ÉTAPE 1 : Choisir des bornes arbitraires de telle sorte que $b^{(0)} < b^{(1)} < \dots < b^{(L-1)} < b^{(L)}$.

ÉTAPE 2 : Calculer la proportion W_h , la moyenne \bar{Y}_h et la variance S_h^2 (à l'aide des équations (3.12), (3.13) et (3.14) respectivement) fondées sur ces bornes, $h = 1, \dots, L-1$.

ÉTAPE 3 : Remplacer la série de bornes initiale par $b^{(1)}, \dots, b^{(L-1)}$ où

$$b^{(h)} = \frac{-\alpha_h' + \sqrt{\beta_h'^2 - 4\alpha_h'\gamma_h}}{2\alpha_h'}, h = 1, \dots, L-1.$$

ÉTAPE 4 : Reprendre les étapes 2 et 3 jusqu'à ce qu'on obtienne deux séries consécutives identiques ou très semblables, c'est-à-dire

$$L-1 \max |b^{(h)} - b^{(h)}'| < \epsilon \text{ pour une valeur } \epsilon < 0.$$

Dans la section qui suit, nous comparons l'application de cet algorithme à celle d'autres méthodes.

4. EXEMPLES

Afin d'illustrer l'algorithme décrit dans la section 3, nous allons nous servir de données tirées des enquêtes annuelles sur le commerce de détail et le commerce de gros réalisées par Statistique Canada. Ces enquêtes permettent d'évaluer le chiffre des ventes des entreprises spécialisées dans le commerce de détail ou le commerce de gros. Nous utilisons trois populations pour illustrer l'algorithme. Il s'agit des grossistes en produits divers du Québec (population 1), des grossistes en produits alimentaires divers du Manitoba (population 2) et des magasins de vente au détail d'appareils ménagers et d'appareils électroniques du Québec (population 3). Le choix des populations s'est fait de manière à obtenir un éventail des tailles de population: grande, moyenne et petite. Les coefficients d'asymétrie pour ces populations sont 24.2 (population 1), 6.5 (population 2) et 13.6 (population 3).

Nous allons comparer les résultats de l'algorithme proposé à ceux de deux autres méthodes. La première de ces méthodes consiste simplement à stratifier la population au moyen de la règle de la racine de f cumulative de Dalenius — Hodges (1959). La seconde méthode consiste à déterminer l'emplacement de la coupure entre la strate à tirage complet et la strate à tirage partiel au

$$B = \sum_{h=1}^H (W_h \sigma_h)^2 (W_h \mu_h)^{-p},$$

$$F = N c^2 \mu^2 + \sum_{h=1}^H W_h \sigma_h^2,$$

$$K_h = B P (W_h \mu_h)^{p-1} - A P (W_h \sigma_h)^2 (W_h \mu_h)^{-p-1},$$

$$T_h = A W_h (W_h \mu_h)^{-p}.$$

Si nous identifions le coefficient de $b^{(h)}$ par α_h , le coefficient de $b^{(h)}$ par β_h et l'ensemble des autres termes par γ_h , nous pouvons représenter les équations (3.8) et (3.9) comme des équations quadratiques de la forme $\alpha_h b^{(h)} + \beta_h b^{(h)} + \gamma_h = 0$. Or, comme le souligne Sethi (1963), les termes α_h, β_h et γ_h sont eux-mêmes des fonctions de $b^{(1)}, \dots, b^{(L-1)}$ par les intégrales (3.1), (3.2) et (3.3). En adoptant l'approche de Sethi (1963), nous pouvons résoudre facilement les équations (3.8) et (3.9) à l'aide de la méthode itérative suivante:

ETAPE 1 : Prendre des bornes arbitraires $b^{(1)} < \dots < b^{(L-1)}$.

ETAPE 2 : Calculer les proportions W_h , les moyennes μ_h et les variances σ_h^2 (à l'aide des équations (3.1), (3.2) et (3.3), respectivement) fondées sur ces bornes, $h = 1, \dots, L-1$.

ETAPE 3 : Remplacer la série de bornes initiale par $b^{(1)}, \dots, b^{(L-1)}$ où

$$b^{(h)} = \frac{-\alpha_h + \sqrt{\beta_h^2 - 4 \alpha_h \gamma_h}}{2 \alpha_h}, h = 1, \dots, L-1. \tag{3.10}$$

ETAPE 4 : Reprendre les étapes 2 et 3 jusqu'à ce qu'on obtienne deux séries consécutives identiques ou très semblables, c'est-à-dire

$$\max |b^{(h)} - b^{(h)}| < \epsilon \text{ pour une valeur } \epsilon > 0. \tag{3.11}$$

Soulignons qu'il est possible de démontrer que le signe précédant la racine carrée ($\sqrt{\quad}$) est positif parce que $b^{(h)}$ se situe entre μ_h et μ_{h+1} . L'inconvénient de cet algorithme est que son utilisation exige que l'on dispose de renseignements sur la fonction de densité approximative $f^{(y)}$. Comme la population considérée est finie, nous pouvons éliminer cet inconvénient en remplaçant les quantités (3.1), (3.2) et (3.3) par les expressions correspondantes fondées sur une population finie. Ainsi, conformément au procédé appliqué dans Cochran (1977), nous pouvons remplacer les paramètres de population infinie donnés par les expressions (3.1), (3.2) et (3.3) par les paramètres correspondants pour population finie, c'est-à-dire:

$$W_h = \frac{N_h}{N}, \tag{3.12}$$

Pour ce qui a trait à la répartition proportionnelle à Y à la puissance p , nous avons l'équation suivante:

$$(3.6) \quad a_h = \frac{\sum_{l=1}^L (W_h^h)^p}{(W_h^h)^p},$$

où $0 < p < \infty$. Dans cet article, nous considérerons surtout la répartition proportionnelle à Y mais les calculs peuvent être effectués aussi bien pour la répartition proportionnelle à N et, de fait, pour n'importe quel mode de répartition représenté par le terme a_h où $\sum_{l=1}^L a_h = 1$. Si nous remplaçons a_h dans l'équation (3.4) par l'expression définie en (3.6), nous obtenons

$$(3.7) \quad n = N W_L + \frac{N \left[\sum_{l=1}^L (W_h^h)^2 (W_h^h)^p \right] \left[\sum_{l=1}^L (W_h^h)^p \right]}{\sum_{l=1}^L W_h^h \sigma_h^2 + \sum_{l=1}^L W_h^h \sigma_h^2}.$$

Afin de déterminer les bornes optimales $b^{(1)}, \dots, b^{(L-1)}$ de manière que la taille n de l'échantillon soit minimale, on calcule les dérivées partielles de l'équation (3.7) par rapport à $b^{(1)}, \dots, b^{(L-1)}$, respectivement et on les pose égales à zéro. On obtient alors les équations suivantes:

$$(3.8) \quad \begin{aligned} & [F T_h - F T_{h+1}] b^{(h)} + \\ & [F K_h - 2\mu_h F T_h - F K_{h+1} + 2\mu_{h+1} F T_{h+1} + 2\mu_h AB - 2\mu_{h+1} AB] b^{(h)} + \\ & [F T_h \mu_h^2 + F T_h \sigma_h^2 - F T_{h+1} \mu_{h+1}^2 - F T_{h+1} \sigma_{h+1}^2 - AB \mu_h^2 + AB \mu_{h+1}^2] = 0, \end{aligned}$$

et pour $h = L-1$,

$$(3.9) \quad \begin{aligned} & [F T_{L-1} - AB] b^{(L-1)} + \\ & [F K_{L-1} - 2\mu_{L-1} F T_{L-1} + 2\mu_L AB] b^{(L-1)} + \\ & [F T_{L-1} \mu_{L-1}^2 + F T_{L-1} \sigma_{L-1}^2 - AB \mu_{L-1}^2 - F T_L \mu_L^2 - F T_L \sigma_L^2] = 0, \end{aligned}$$

où

$$A = \sum_{l=1}^L (W_h^h)^p,$$

Le problème consiste à déterminer des bornes $b^{(1)}, b^{(2)}, \dots, b^{(L-1)}$ (où $y^{(1)} < b^{(1)} < \dots < b^{(L-1)} < y^{(N)}$) de manière à minimiser la taille n de l'échantillon global, étant donné le degré de fiabilité c et le mode de répartition spécifiée (représentée par q_h).

3. L'ALGORITHME

La méthode utilisée dans le présent article pour calculer des bornes de strates pour un degré de précision voulu a déjà été utilisée par Dalenius (1950) pour calculer des bornes de strates pour une taille d'échantillon donnée. On suppose en premier lieu que l'échantillonnage se fait à partir d'une population dont la distribution de fréquences peut être représentée avec suffisamment de précision par une fonction continue de densité $f(y)$. Ainsi, pour un ensemble donné de limites $b^{(1)}, \dots, b^{(L-1)}$ nous définissons les quantités suivantes:

$$W_h = \int_{b^{(h)}}^{b^{(h+1)}} f(y) dy, \tag{3.1}$$

$$\mu_h = \int_{b^{(h)}}^{b^{(h+1)}} y f(y) dy / W_h, \tag{3.2}$$

$$\sigma_h^2 = \int_{b^{(h)}}^{b^{(h+1)}} y^2 f(y) dy / W_h - \mu_h^2, \tag{3.3}$$

pour $h = 1, \dots, L$, où $b^{(o)} = -\infty, b^{(L)} = +\infty$.
On peut alors reformuler l'équation (2.4) comme suit:

$$n = NW_L + \frac{N^2 \sum_{h=1}^{L-1} W_h^2 \sigma_h^2 / a_h}{N^2 \sum_{h=1}^{L-1} W_h^2 \sigma_h^2 + N^2 \mu_h^2}, \tag{3.4}$$

où

$$\mu = \int_{b^{(L)}}^{b^{(o)}} y f(y) dy.$$

Il convient de souligner que même s'il s'agit d'une population de grande taille, le facteur de correction pour population finie (cpt) est toujours présent dans l'équation (3.4) - voir Dalenius-Gurney (1951). Par définition, la strate à tirage complet doit avoir une population finie si l'on veut que la taille d'échantillon soit finie. En outre, le fait d'ignorer le facteur de c.p.f. ne se traduirait pas par une variance nulle pour la strate à tirage complet.
Le terme " q_h " dans l'équation (2.3) peut aussi être exprimé à l'aide des quantités (3.1), (3.2) et (3.3). Dans le cas de la répartition proportionnelle à N à la puissance p , nous avons:

$$a_h = \frac{W_h^p}{\sum_{h=1}^L W_h^p}, \tag{3.5}$$

pour $h = 1, \dots, L-1$.

où $y_{M^{h-1}+1} \leq z_j \leq y_{M^h}$ pour $j = m_{h-1}+1, \dots, m_h$ ($h = 1, 2, \dots, L-1$), $m_h = \sum_{i=1}^I n_i$ pour $h = 1, 2, \dots, L$ et $m_0 = 0$.

Supposons que le degré de précision voulu pour la moyenne estimée est défini par c (coefficient de variation) et que la proportion d'unités échantillonnées devant être réparties entre les $L-1$ premières strates est a_h ($h = 1, 2, \dots, L-1$) où $\sum_{h=1}^{L-1} a_h = 1$. Le terme " a_h " représente en fait n'importe quel mode de répartition entre les strates. Dans le cas, par exemple, de la répartition proportionnelle à N à la puissance p ,

$$a_h = \frac{N_h^{p-1}}{\sum_{h=1}^{L-1} N_h^{p-1}} \quad (h = 1, 2, \dots, L-1)$$

et dans le cas de la répartition proportionnelle à Y à la puissance p ,

$$a_h = \frac{Y_P^h}{\sum_{h=1}^{L-1} Y_P^h},$$

où $0 < p < \infty$. Etant donné des hypothèses relativement simples et une valeur de p appropriée, la répartition à la puissance p a la propriété d'engendrer des coefficients de variation relative-ment uniformes pour les strates à tirage partiel sans accroître notablement le coefficient de variation global. Cette uniformité des coefficients de variation est souvent souhaitée par les utilisateurs des données d'enquête.

En pratique, on choisit souvent $1/2$ ou $1/3$ comme valeur de p . Une faible valeur de p (c.-à-d., une valeur proche de 0) donne habituellement des coefficients de variation comparables d'une strate à l'autre tandis qu'une valeur plus élevée élargit l'écart entre les coefficients mais accroît du même coup la précision des estimations globales.

Souignons également que la répartition à la puissance p définie ci-dessus équivaut au mode proposé par Bankier (1988) lorsque les coefficients de variation des strates à tirage partiel sont égaux.

La variance de \hat{Y} est

$$V(\hat{Y}) = \frac{1}{N^2} \sum_{h=1}^{L-1} n_h (N_h - n_h) S_h^2 \quad (2.3)$$

où S_h^2 désigne la variance de population de chaque strate h . Si l'on tient compte du degré de précision voulu (coefficient de variation c), la variance de Y peut être redéfinie $V(Y) = c^2 Y^2$. En remplaçant n_h et $V(Y)$ dans l'équation (2.3) par $(n - N^L) a_h$ et $c^2 Y^2$ respectivement et en isolant n , on obtient l'équation

$$n = N^L + \frac{\sum_{h=1}^{L-1} N_h^2 S_h^2 / a_h}{(N c Y)^2 + \sum_{h=1}^{L-1} N_h S_h^2} \quad (2.4)$$

visé à recueillir des données sur les ventes mensuelles. Pour ce qui a trait aux enquêtes polyvalentes, si l'on utilise des variables auxiliaires qui ne sont pas en corrélation étroite avec la variable principale, les bornes de strates ne seront plus aussi optimales. L'algorithme que nous proposons est une version modifiée de la méthode de stratification de Sethi (1963). Les bornes de strates qui découlent de cet algorithme sont optimales et déterminent la taille d'échantillon minimum requise. Le mode de répartition que nous avons choisi pour illustrer la méthode est la répartition à la puissance p . L'utilisation de ce mode de répartition permet de publier des estimations de strates dont les coefficients de variation ne sont pas très différents les uns des autres. La répartition à la puissance p a été proposée par Carroll (1970), Fellegi (1981) et Bankier (1988). En pratique, elle est un compromis entre la répartition de Neyman et la nécessité d'avoir des coefficients de variation égaux pour chaque strate. Un inconvénient de la répartition de Neyman est qu'elle produit des coefficients de variation très différents d'une strate à l'autre si l'on doit avoir des estimations pour chaque strate. Par ailleurs, un mode de répartition qui produit les mêmes coefficients de variation d'une strate à l'autre peut exiger un échantillon beaucoup plus grand que celui qu'exige la répartition de Neyman. Dans le cadre de notre analyse, la répartition à la puissance p nous permettra de produire des estimations pour des strates d'entreprises de tailles variées (petites, moyennes et grandes) avec des coefficients de variation similaires.

Nous comparerons la méthode élaborée dans cet article, au point de vue des bornes et de la taille d'échantillon, à la règle de la racine de f cumulative de Dalenius — Hodges (1959) ainsi qu'à une combinaison des méthodes de stratification d'Hidiroglou (1986) et de Dalenius — Hodges (1959). L'algorithme, qui est récursif par définition, est facile à programmer et converge rapidement vers les bornes optimales. Il permet également une réduction sensible de la taille d'échantillon pour des critères de fiabilité donnés.

2. LE PROBLÈME

Considérons une population ordonnée finie de N unités:

$$Y^{(1)}, Y^{(2)}, \dots, Y^{(N)},$$

où $Y^{(i)} \leq Y^{(i+1)}$ pour $i = 1, 2, \dots, N-1$. Cette population est découpée en L strates. Le nombre d'unités dans chaque strate est désigné par N_h , $h = 1, 2, \dots, L$. Selon le plan d'échantillonnage, n_h unités doivent être prélevées dans chaque strate à tirage partiel de taille N_h ($h = 1, 2, \dots, L-1$) au moyen d'un échantillonnage aléatoire simple sans remise; notons par ailleurs que $n_L = N_L$. La moyenne à estimer est

$$\bar{Y} = \frac{\sum_{L=1}^L \sum_{h=1}^{M_h} Y^{(j)} / N_h}{\sum_{h=1}^L M_h} \quad (2.1)$$

où $M_h = \sum_{i=1}^I N_i$ pour $h = 1, 2, \dots, L$ et $M_0 = 0$.

Compte tenu de cette définition, l'estimateur de la moyenne de population \bar{Y} est

$$\hat{\bar{Y}} = \frac{\sum_{L=1}^L \sum_{h=1}^{M_h} n_h \bar{Y}_h + \sum_{j=1}^{M_{L-1}+1} Z_j}{\sum_{N=1}^{M_L-1+1} Y^{(j)}} \quad (2.2)$$

Sur la stratification de populations asymétriques

PIERRE LAVALLÉE et MICHEL A. HIDIROGLOU¹

RÉSUMÉ

Pour un degré de précision donné, Hidiroglou (1986) a défini un algorithme permettant de diviser la population en une strate à tirage complet et en une strate à tirage partiel de manière à minimiser la taille de l'échantillon global en supposant un échantillonnage aléatoire simple sans remise dans la strate à tirage partiel. Sethi (1963) a proposé un algorithme permettant un découpage optimal de la population en un certain nombre de strates à tirage partiel. Dans cet article, il est question d'un algorithme itératif qui vise à déterminer les bornes de strates pour une population fortement asymétrique découpée en une strate à tirage complet et en un certain nombre de strates à tirage partiel. Ces bornes de strates sont calculées de manière à minimiser la taille de l'échantillon global étant donné un degré de précision relative, un échantillonnage aléatoire simple sans remise dans les strates à tirage partiel et une répartition à la puissance "p" de l'échantillon entre ces mêmes strates. L'algorithme présenté dans cet article est une combinaison des travaux d'Hidiroglou (1986) et de Sethi (1963).

MOTS CLÉS: Algorithme itératif; bornes optimales; tirage complet; tirage partiel.

1. INTRODUCTION

Dans le cas de populations fortement asymétriques comme celles que l'on retrouve dans les enquêtes entreprises, un échantillonnage efficace n'est possible qu'à condition de découper ces populations en une strate à tirage complet et en un certain nombre de strates à tirage partiel. Toutes les unités de la strate à tirage complet sont prélevées avec une probabilité égale à un (1) tandis que les unités des strates à tirage partiel sont prélevées suivant un modèle probabiliste. Glasser (1962) et Hidiroglou (1986) ont établi des règles approximatives pour le découpage d'une population en une strate à tirage complet et en une strate à tirage partiel. Glasser (1962) a déterminé l'emplacement de la coupure entre les strates en supposant qu'un échantillon de taille déterminée devait être prélevé dans les deux strates et que dans le cas de la strate à tirage partiel, il devait s'agir d'un échantillonnage aléatoire simple sans remise. Pour sa part, Hidiroglou (1986) a déterminé l'emplacement de la coupure entre les strates en supposant qu'il fallait respecter un degré de précision voulu. Ces deux approches sont duales en ce sens que Glasser cherche à minimiser la variance d'échantillonnage pour une taille d'échantillon donnée tandis que Hidiroglou cherche à minimiser la taille d'échantillon pour une variance d'échantillonnage donnée.

Dans cet article, nous proposons un algorithme qui vise à découper une population fortement asymétrique en une strate à tirage complet et en un certain nombre de strates à tirage partiel. Par cet algorithme, nous cherchons à minimiser la taille de l'échantillon global étant donné le coefficient de variation de l'estimateur et le mode de répartition de l'échantillon entre les strates à tirage partiel. Les bornes de strates sont déterminées en fonction d'une variable auxiliaire qui est en corrélation étroite avec les renseignements recueillis dans l'enquête. Si, par exemple, le chiffre des ventes annuelles est une des variables mesurées dans un recensement des détaillants, cette variable auxiliaire peut servir à déterminer les bornes de strates pour une enquête spécialisée qui

¹ Pierre Lavallée, méthodologiste, et Michel A. Hidiroglou, chef, Division des méthodes d'enquêtes-entreprises, Statistique Canada, Ottawa (Ontario) K1A 0T6, Canada. Les auteurs tiennent à remercier France Blocq de la Division des méthodes d'enquêtes-entreprises de Statistique Canada pour avoir programmé les exemples.

- KING, A.J., et JESSEN, R.J. (1945). The master sample of agriculture, *Journal of the American Statistical Association*, 38-56.
- KISH, L. (1987). *Statistical Design for Research*: New York: Wiley-Interscience.
- KISH, L. (1986). Timing of surveys for public policy. *Australian Journal of Statistics*, 28, 1-12.
- KISH, L. (1980). Design and estimation for domains. *The Statistician*, London, 29, 209-222.
- KISH, L. (1976). Optima and proxima in linear sample designs. *Journal of the Royal Statistical Society*, A, 139, 80-95.
- KISH, L. (1965). *Survey Sampling*. New York: John Wiley and Sons.
- KISH, L. (1961). Efficient allocation for multipurpose samples. *Econometrica*, 29, 363-385.
- KISH, L., et ANDERSON, D.W. (1978). Multivariate and multipurpose stratification, *Journal of the American Statistical Association*, 73, 24-34.
- KISH, L., et FRANKEL, M.R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society*, B, 36, 1-37.
- KISH, L., et SCOTT, A.J. (1971). Retaining units after changing strata and probabilities, *Journal of the Royal Statistical Society*, 66, 461-470.
- KIREGYERA, B., et GACHUKI, P. (1985). Experiences in panel surveys: examples from an integrated sample survey programme in Kenya. *Bulletin of the International Statistical Institute*.
- MACURA, M., et CLELAND, J. (1985). *A Celebration of Statistics: the ISI Centenary Volume*, (éds. A.C. Atkinson et S.E. Fienberg), New York: Springer Verlag.
- MURTHY, M.N. (1974). Evaluation of multi-subject sample survey systems. *International Statistical Review*, 42.
- MURTHY, M.N. (1967). *Sampling Theory and Methods*. Calcutta: Statistical Publishing Society.
- NATIONS-UNIES (1980). *National Household Survey Capability Programme* (NHSCP). New York: United Nations.
- RODRIGUEZ-VERA, A. (1982). *Multipurpose Optimal Sample Allocation Using Mathematical Programming*. Thèse de doctorat, The University of Michigan, Ann Arbor.
- VERMA, V., SCOTT, C., et O'MURCHEARTAIGH, C. (1980). Sample designs and sampling errors for the World Fertility Survey. *Journal of the Royal Statistical Society*, A, 143, 431-473.
- VATES, F. (1981). *Sampling Methods for Censuses and Surveys*, (4^e éd.). London: Griffin and Co.

À cette titre, les méthodes de calcul et de présentation des erreurs d'échantillonnage méritent d'être mises en évidence parmi les nombreuses statistiques produites habituellement à la suite d'enquêtes. Il ne suffit pas de présenter l'erreur type d'un seul ou de quelques-uns des paramètres statistiques les plus importants: ils sont trop nombreux et trop variés. C'est pourquoi on en est venu à calculer à l'aide des variances d'autres expressions de la variabilité d'échantillonnage, notamment les estimations des "effets du plan" d_g^2 et parfois, à l'aide de $d_g^2 = 1 + \rho_g(d_g^2 - 1)$, les estimations des coefficients de corrélation intraclass synthétique ρ_g .

En résumé, voici les points à retenir: a) calculer des erreurs d'échantillonnage pour de nombreuses variables parce que les variances, les effets du plan (d_g^2) et les coefficients de corrélation intraclass (ρ_g) varient énormément d'une variable à l'autre; b) vous devrez peut-être faire des moyennes d'erreurs d'échantillonnage car il pourrait être inopportun ou embarrassant de les présenter toutes; c) il pourrait être ni possible ni nécessaire de calculer des erreurs d'échantillonnage pour toutes les sous-classes car on peut souvent obtenir une approximation de ces erreurs à l'aide de modèles acceptables; d) il est nécessaire de présenter les erreurs d'échantillonnage pour les sous-classes et d'autres paramètres statistiques afin de guider les personnes qui prennent connaissance des rapports (Kish 1965; Kish 1987; Verma et coll. 1980). Nous espérons que ce sujet recevra à l'avenir, de la part des théoriciens et des méthodologistes, toute l'attention qu'il mérite.

10. CONCLUSIONS

Dans les sections 4 à 9, nous avons proposé des approches et des solutions très diversifiées pour le calcul d'une moyenne des résultats de répartition de la taille de l'échantillon entre les domaines semble produire des solutions de compromis étonnamment bonnes. L'utilisation d'un plus grand nombre de critères de stratification, recommandée dans la section 6, peut aussi procurer des avantages appréciables. Dans les sections 4 et 7, des considérations relatives aux estimations de sous-classe ont abouti à des décisions totalement différentes en ce qui concerne les plans de sondage. Dans la section 8, nous avons vu comment adapter le mieux possible les plans de sondage pour enquêtes périodiques aux objectifs de ces enquêtes et comment ces plans constituent le meilleur compromis pour les enquêtes à usages multiples. Nous avons examiné simultanément les divers niveaux d'usages et les diverses sources d'incompatibilité. Dans tout problème, l'essentiel est de savoir poser la bonne question. Nous proposons donc le plan à usages multiples comme un nouveau modèle destiné à remplacer les solutions "optimales" à des questions arbitrairement partielles comme la suivante: Quelle est la répartition optimale pour la moyenne Y ou le total Y de la variable "la plus importante"?

BIBLIOGRAPHIE

- BEAN, J. A., et BURMEISTER, L. F. (1978). A review of optimal sample allocation for multipurpose surveys, *Biometrika*, 20, 3-14.
- CHATTERJEE, S. (1967). A note on optimum stratification, *Skandinavisks Actuarietidskrift*, 50, 40-44.
- COCHRAN, W. G. (1977). *Sampling Techniques*, (3^e éd.). New York: John Wiley and Sons, Inc.
- DALENIUS, T. (1957). *Sampling in Sweden*, Stockholm: Almqvist and Wiksell.
- DUNCAN, G. J., et KALTON, G. (1986). Issues of design and analysis of surveys across time. *International Statistical Review*, 54.
- FELLEGI, I. P., et SUNTER, A. B. (1974). Balance between different sources of survey errors. *Sankhyā*, 36, 119-142.
- FOREMAN, E. K. (1983). Integrated programmes of household surveys: design aspects. *Bulletin of the International Statistical Institute*.

Tableau 4
Usages et plans de sondage des enquêtes périodiques

Usages	Plans	Mode de renouvellement
A. Niveaux courants	A. Chevauchement partiel	
B. Valeurs cumulatives	0 < P < 1	abc-cde-efg
C. Variations nettes (moyennes)	B. Aucun chevauchement P = 0	aaa-bbb-ccc
D. Variations brutes	C. Chevauchement complet P = 1	aaa-aaa-aaa
(valeurs individuelles)	D. Panels	mêmes éléments
E. Séries chronologiques à usages multiples	E. Combinaisons, plan à panel	
	F. Bases principales	

ou des coûts. Par ailleurs, le calcul des variations individuelles (variations brutes ou micro-variations) (D) nécessite des panels tandis que le calcul des valeurs cumulatives (B) nécessite une modification des échantillons et est plus rapide lorsqu'il n'y a pas de chevauchement. En ce qui concerne les niveaux courants (A), il est possible de réduire quelque peu les variances à l'aide d'estimateurs qui exploitent des corrélations découplant de chevauchements partiels. Dans le cas des variations nettes (C), l'estimation profite des corrélations découplant de n'importe quel degré de chevauchement et dans la plupart des autres cas, l'estimation profite des corrélations qui découlent de chevauchements complets (Cochran 1977; Kish 1987; Kish 1965). On peut souvent en arriver à des compromis acceptables lorsqu'on peut définir les usages. Cependant, des considérations accessoires peuvent écartier certains plans (par exemple, les chevauchements peuvent être défendus ou permis) et imposer l'utilisation de plans moins efficaces mais tout aussi acceptables.

Les six plans de sondage du tableau 4 se distinguent surtout par le degré (et le genre) de chevauchement entre les périodes. Dans le cas du chevauchement complet, le mode de renouvellement (aaa-aaa) indique que les portions sont les mêmes pour toutes les périodes; dans le cas où il n'y a aucun chevauchement (aaa-bbb), le mode de renouvellement n'indique aucune portion commune et dans le cas du chevauchement partiel (abc-cde-efg), c et e ne représentent qu'un chevauchement d'un tiers entre des périodes successives. Nous nous intéressons ici plus particulièrement aux effets sur différents usages de la variation du degré de chevauchement P dans divers plans (chevauchement complet: P = 1; aucun chevauchement: P = 0 et chevauchement partiel: 0 < P < 1). Les usages sont analysés en fonction des variances des moyennes estimées puisque celles-ci (de même que les pourcentages, les taux et les proportions) sont les estimations les plus courantes et les plus simples. Les effets sur les autres estimations ne seront pas radicalement différents mais ils sont trop nombreux, trop variés et trop complexes pour faire l'objet d'une analyse dans la présente étude.

La question des panels (avantages, inconvénients, problèmes et solutions) est traitée plus en détail dans d'autres ouvrages (Duncan et Kalton 1986; Kish 1987). Nous attirons plus particulièrement l'attention du lecteur sur les PPF, ou plans à panel fractionné, dont nous encourageons l'utilisation comme plans de sondage à usages multiples. Le PPF combine un échantillon constant P avec de nouveaux échantillons avec renouvellement de sorte que Pa-Pb-Pc-Pd désignent les échantillons périodiques. Les échantillons avec renouvellement a, b, c, d, etc. peuvent être groupés de manière à former de plus gros échantillons. L'échantillon constant P sert principalement à produire des estimations sur les micro-variations (variations brutes individuelles) mais il assure aussi le chevauchement partiel nécessaire à de meilleures estimations des niveaux courants et des macro-variations (variations nettes — moyennes) pour n'importe quelle paire de périodes.

9. CALCUL ET PRÉSENTATION DES ERREURS D'ÉCHANTILLONNAGE

Bien que l'examen de cette question dans le cadre de la présente étude semble discutable, nous sommes convaincus qu'il s'agit là d'un problème qui touche les plans de sondage à usages multiples.

particulièrement des compromis rigoureux. Il est nécessaire de distinguer ces deux niveaux d'usages (tableau 1) parce que les enquêtes portant sur plusieurs sujets utilisent chacune un seul échantillon dans une opération tandis que les opérations d'enquête intégrées peuvent utiliser des unités de sondage de tailles différentes pour des enquêtes industrielles ou commerciales: les mesures de la taille par exemple des plans intégrés pour les populations totales et l'agriculture, peut-être aussi pour les sous-populations ethniques et les activités industrielles ou commerciales: les mesures de la taille pour chacun de ces cas peuvent différer largement l'une de l'autre. Il est néanmoins possible de trouver un solution de compromis de manière à obtenir un niveau d'efficacité raisonnable dans chaque cas.

Les mesures de la taille ont également un rapport étroit avec les problèmes qui se rattachent à la conservation des unités par suite d'une modification des strates et des probabilités (Kish et Scott 1971). Ces méthodes ont été conçues pour tenir compte de l'évolution des unités de sondage tant au point de vue des mesures de la taille qu'au point de vue des variables de stratification, mais elles peuvent aussi tenir compte des différences entre les variables d'enquête:

“On attribue souvent des probabilités d'échantillonnage inégales aux unités de sondage. Bien qu'elles aient une application plus vaste, nos méthodes servent spécialement à choisir des unités primaires d'échantillonnage pour les enquêtes. Le plus souvent, ces unités sont prélevées séparément dans de nombreuses strates à raison d'une unité par strate.”

“Après l'échantillonnage initial, les unités peuvent servir à de nombreuses enquêtes pendant plusieurs années mais progressivement, les besoins des nouvelles enquêtes seront probablement mieux servis par des strates et des probabilités d'échantillonnage fondées sur de nouvelles données que par les strates et les probabilités qui ont servi à l'échantillonnage initial. La différence entre les données initiales et les nouvelles données peut être imputable à des changements particuliers dans les unités de sondage, rapportées par le dernier recensement ou à une modification des objectifs et des populations de l'enquête. Par exemple, un échantillon conçu à l'origine pour les ménages et les personnes peut devoir servir ultérieurement à une enquête sur les agriculteurs ou les étudiants de niveau collégial. Evidemment, nos méthodes peuvent aussi servir à constituer un groupe d'échantillons connexes qui n'ont pas les mêmes objectifs.”

La méthode proposée permet d'appliquer les meilleures mesures (pour la taille et les strates) pour chaque usage de l'échantillon tout en conservant le plus grand nombre possible d'unités de sondage (surtout des UPE) communes à plusieurs échantillons pour des usages différents. Il serait toujours possible d'imaginer un compromis par lequel on ferait la moyenne des mesures de manière à obtenir un chevauchement complet des unités même si cela devait impliquer une perte d'efficacité pour chacun des usages. Une combinaison des deux méthodes peut s'avérer encore plus efficace, soit accroître le degré de chevauchement avec une faible perte d'efficacité dans chaque cas en considérant uniquement les différences de mesures qui excèdent une valeur minimum arbitraire (Kish et Scott 1971).

8. USAGES ET PLANS DE SONDAGE DES ENQUÊTES PÉRIODIQUES

Les enquêtes périodiques soulèvent des difficultés de plus en plus grandes à mesure que leur nombre et leur importance s'accroissent. Il est faux de supposer que ces enquêtes coûteuses et déterminantes servent à un seul des cinq usages énumérés dans le tableau 4 puisqu'elles sont destinées habituellement à tous ces usages ou à plusieurs d'entre eux si le plan de sondage le permet. Le tableau 4 contient cinq usages et six plans. Les quatre premiers éléments de chaque groupe sont désignés par des lettres identiques d'une ligne à l'autre. Cette correspondance signale les plans de sondage qui conviennent le mieux (par des variances réduites) à chacun des quatre usages. Comme la plupart des enquêtes périodiques poursuivent plusieurs objectifs, il est normal que nous nous attaquions au délicat problème des plans de sondage à usages multiples. Dans la réalité, n'importe quel des six plans énumérés dans le tableau 4 peut servir à calculer les niveaux courants (A) et les variations nettes (C) mais cela se fait difficilement sans une augmentation de la variance

variables d'enquête différentes tendront à avoir des relations optimales variées avec les critères de stratification; il est donc fortement recommandé d'utiliser un grand nombre de critères même si chacun d'eux n'est utilisé qu'avec quelques divisions de strates (catégories). Le plan de sondage à usages multiples est ce qui se prête le mieux à la stratification à plusieurs variables (Kish et Anderson 1978). C'est peut-être aussi ce qui justifie le mieux le besoin de méthodes "d'échantillonnage contrôlé". Le choix des limites de strates, appelé "stratification optimale", est un sujet connexe mais secondaire dans le cadre du présent article.

7. TAILLES DE GRAPPES; MESURES DE LA TAILLE; CONSERVATION D'UNITÉS

Nous constatons parfois dans la description des plans de sondage que l'effet du plan a été estimé au moyen de la formule $D_g^2 = [1 + \rho_g b'_i - 1]$, où ρ désigne une corrélation intraclass syn-thétique entre la variable "la plus importante", g et $b'_i = n/a$, la taille de grappe moyenne. Il est alors possible de calculer la variance effective des unités élémentaires $S_g^2 D_g^2$ et la variance $S_g^2 D_g^2/n$ pour la moyenne de la variable g . Il faut toutefois s'interroger sur la composition de n et de b_i . Si la population à l'étude est constituée de femmes mariées en âge de procréer, il se peut que cette catégorie de personnes ne représentent que 10% de la population totale et ne se retrouvent que dans 30% des logements et même beaucoup moins en ce qui a trait à quelques rares popula-tions. Cette question a été traitée dans l'échantillonnage pour les caractéristiques peu courantes (Kish 1965). Nous évitons ordinairement les grandes grappes à cause des effets négatifs qu'elles ont sur la variance. Cependant, même de grandes grappes de la population ne produiront que de petites grappes d'individus ayant une caractéristique peu courante, si cette catégorie de per-sonnes sont disséminées un peu partout. Par exemple, on peut échantillonner des îlots complets pour les personnes âgées de plus de 65 ans; on peut aussi sonder des villages complets pour trouver des personnes atteintes d'une maladie bien définie. Par contre, si la caractéristique se retrouve surtout dans de petites régions, il est facile d'identifier ces régions et de les stratifier en conséquence. Dans les plans à usages multiples, les classes de recoupement de l'échantillon ont des tailles variées qui représentent chacune une partie de la taille n_i de l'échantillon et M_g représente la pro-portion de l'effectif de ces classes par rapport à la population. Ainsi, par le plan de sondage nous voulons estimer $[1 + \rho_g(b'_i - 1)]$ non seulement pour diverses variables g pour l'échantillon intégral n_i mais aussi pour de nombreuses classes de recoupement. Comme dans la section précédente, l'indice g sert ici à désigner les variables aussi bien que les sous-classes de manière à simplifier la notation. Nous posons ensuite quelques hypothèses qui se sont avérées de bonnes approximations dans des milliers de calcul empiriques pour un très grand nombre d'échantillons:

$$[1 + \rho_g(b_g^i - 1)] = [1 + \rho_g(M_g^i b'_i - 1)] \approx [1 + \rho_i(M_g^i b'_i - 1)] \quad (2)$$

En d'autres termes, nous utilisons $b_g^i = M_g^i b'_i$ et ρ_g , comme approximations grossières. Il est vrai que cela contribue à sous-estimer quelque peu les valeurs moyennes de D_g^2 pour les classes de recoupement à cause de la variabilité des tailles de grappes. Cependant, cette sous-estimation a des effets négligeables comparativement à ceux qui découlent des fortes variations de ρ_g^i d'une variable à l'autre (Kish 1987; Verma et coll. 1980; Kish et coll. 1976), et influe peu sur l'efficacité des plans de sondage. Il est important de considérer l'efficacité des estimations pour les sous-classes comme pour l'échantillon au complet; cet examen révèle une efficacité beaucoup plus forte pour les grandes grappes que celle que nous aurions observée pour b'_i et n_i si nous en étions tenus à l'échantillon proprement dit. Les mesures de la taille ont un rapport avec la taille des grappes mais varient à cause des erreurs d'entachées les mesures existantes, lesquelles erreurs sont surtout attribuables à une varia-tion de la composition de la population et à l'obsolescence des données. Signalons aussi les pro-blèmes relatifs aux mesures de la taille pour les enquêtes portant sur plusieurs sujets à la fois et les opérations d'enquête intégrées pour des populations différentes; ces problèmes peuvent exiger plus

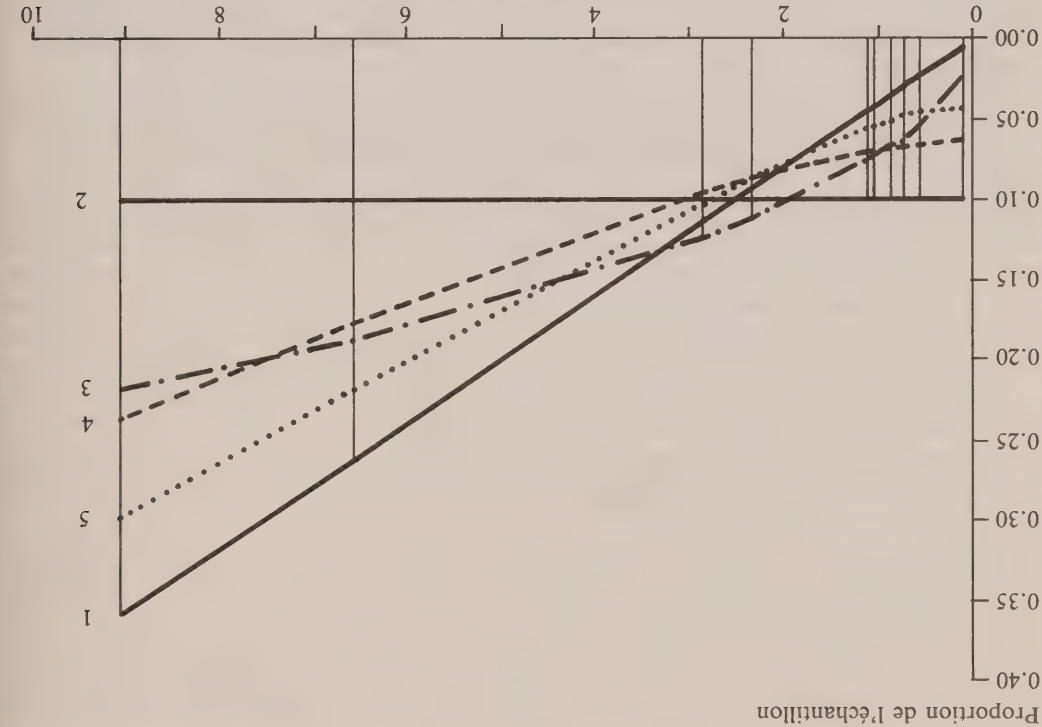


Figure 2. Cinq modes de répartition de la taille n_h de l'échantillon (pour un total fixe $2n_h$)

Les dix provinces du Canada permettent d'illustrer dans le graphique ci-dessus les problèmes qui découlent habituellement de l'existence de grands domaines de tailles inégales et les compromis avantageux qu'il est possible d'établir.

- 1 La répartition proportionnelle à la taille des domaines ($n_h \propto W_h$) est représentée par une diagonale.
- 2 La répartition uniforme ($n_h \propto 1/H$) est représentée par une droite horizontale.
- L'écart entre ce mode de répartition et le précédent est appréciable à l'extrémité des courbes.
- 3 La répartition par la racine carrée ($n_h \propto \sqrt{W_h}$) produit des compromis aux deux extrémités.
- 4 La répartition optimale ($n_h \propto \sqrt{(W_h^2 + 1/H^2)}$) améliore les deux extrémités mais on constate un "aplatissement" intéressant à l'extrémité gauche de la courbe.
- 5 La répartition optimale "pondérée" ($n_h \propto \sqrt{(.8W_h^2 + .2/H^2)}$) améliore considérablement la partie supérieure de la courbe.

populations pour lesquelles on calcule des estimations propres tandis que les strates sont habituellement des divisions de ces sous-populations, qui sont créées pour réduire les variances. Si l'on considère par exemple les provinces comme des domaines, on y créera plus de strates dans le but de réduire les variances à l'échelle de la province; en revanche, les domaines de recoupement comme l'âge, le sexe et la situation économique tendent à chevaucher les strates. Bien qu'elle ne soit pas aussi capitale que la répartition entre les domaines, la répartition de la taille de l'échantillon entre les strates peut être essentielle dans le cas des répartitions optimales efficaces disproportionnées. Les deux méthodes que nous avons décrites dans la Section 5 pour la répartition de la taille de l'échantillon entre les domaines peuvent aussi servir à la répartition entre les strates mais dans des buts différents. Quelques-unes des sources sur la programmation non linéaire concernent les domaines et d'autres les strates tandis que d'autres encore confondent les deux.

La présence de plusieurs variables d'enquête et de plusieurs paramètres statistiques parmi les usages justifie clairement l'utilisation d'un plus grand nombre de variables de stratification. Des

Tableau 3

Fonctions de perte (1 + L) pour deux populations (Kish 1976)

Résultats de répartition m_i	(A)		(B)	
	(1 + L) pour $W_2/W_2 = 4$		(1 + L) pour 133 pays: 0.2 à 100 millions	
	$\Sigma W_i/p_i$	$\Sigma p_i/2$	$\Sigma W_i/p_i$	$\Sigma p_i/133$
M/H	1	1.56	1	6.86
M/H'	1.36	1	3.34	1
$\infty \sqrt{(W_i'}$	1.08	1.125	1.35	1.54
$\infty \sqrt{(W_i^2 + H - 2)}$	1.116	1.080	1.31	1.28
$\infty \sqrt{(0.5W_i^2 + H - 2)}$				1.295
$\infty \sqrt{(2W_i^2 + H - 2)}$			1.47	1.17
$\infty \sqrt{(4W_i^2 + H - 2)}$			1.20	1.44
			1.12	1.66
				(1.39)
				1.23

Dans le cas (A), il y a deux strates et deux domaines ($W_1 = 0.8$ et $W_2 = 0.2$); il convient de souligner que $m_i = \sqrt{W_i}$ produit un résultat presque aussi bon que la répartition optimale dans le cas de la fonction de perte combinée. Dans le cas (B), nous considérons les populations de 133 pays, dont la taille varie de 200,000 à plus de 100,000,000, soit un rapport de 1 à 500. Pour des raisons pratiques, nous avons exclu de notre analyse (par rapport au traitement qui a été fait pour l'enquête mondiale sur la fécondité) les quatre plus importants pays et les quelques pays qui comptent moins de 200,000 habitants. Leur inclusion aurait fait passer la variance des rapports de tailles, W_i , de 2.5 à 12 et aurait produit des résultats spectaculaires. Il convient de souligner que la répartition $\sqrt{W_i}$ réduit sensiblement les pertes. Un compromis est préférable à rien du tout. Néanmoins, la répartition optimale, $\sqrt{(W_i^2 + H - 2)}$, est bien supérieure. Diverses valeurs de $I_c/I_p (= 1/2, 2/1$ et $4/1)$ augmentent légèrement la variance de la fonction de perte combinée avec des poids (1:1) mais produisent des résultats stables pour les fonctions de perte combinées avec poids ($I_c/I_p:1$). Deux exemples dans le tableau 3 illustrent les compromis étonnamment satisfaisants qui ont pu être établis entre des résultats de répartition contradictoires grâce à la méthode de la moyenne pondérée: les données pertinentes, qui se trouvent à la quatrième ligne du tableau 3, se comparent très avantageusement aux autres données du tableau. Ces résultats s'expliquent par l'existence d'une très longue courbe horizontale pour la répartition optimale, comme nous l'avons souligné dans la Section 2 et dans d'autres ouvrages (Kish 1976; Kish 1987). Au Canada, par exemple, le rapport de tailles de population entre les provinces peut atteindre 1/70, ce qui nous amène à dire que le cas des dix provinces canadiennes peut être apparenté au cas B du tableau 3; par ailleurs, le graphique de la figure 2 a été construit à l'aide des chiffres de population des dix provinces. (Vois aussi Fellegi et Sunter 1974.)

le coût total fixe $\Sigma C_i n_i$. En ce qui a trait à la controverse entourant la définition de n_d ($n_d = n/H$ — échantillons de même taille pour les domaines ou $n_d = nW_d$ — taille d'échantillon proportionnelle à l'effectif W_d du domaine), nous pouvons affirmer que les résultats de la répartition optimale (de compromis) sont proportionnels à $\sqrt{(W_d^2 + H - 2)}$, avec des poids identiques I_i . Les enquêtes mondiales sur la fécondité illustrent bien le cas qui nous occupe; dans ces enquêtes, on utilise des échantillons de taille comparable pour les petits et les grands pays. Les tailles d'échantillon se situaient uniquement entre 3 et 10 000, sans corrélation apparente avec la taille de la population. Les variances des moyennes continentales des enquêtes nationales ont par conséquent doublé ou triplé. On décrit dans ces termes l'apport des enquêtes sur la fécondité: "Le principal apport de ces enquêtes a été jusqu'à maintenant de confirmer la baisse soutenue de la fécondité, qui a caractérisé une bonne partie de l'Asie et de l'Amérique latine dans les années 1970, et de faire ressortir le contraste entre ces continents et l'Afrique, où le taux de fécondité demeure élevé et où la volonté d'avoir de nombreux enfants est encore très réelle" (Macura et Cleland 1985). (TRADUCTION).

6. RÉPARTITION ENTRE LES STRATES ET CHOIX DES CRITÈRES

DE STRATIFICATION

On confond souvent les domaines et les strates dans les discussions mais il importe de bien les distinguer dans les études pratiques sur les plans de sondage. Les domaines sont des sous-

Burmeister 1978; Rodriguez-Vera 1982 et Cochran 1977). Le "coût minimum requis" est souvent beaucoup trop élevé parce que les exigences de fiabilité sont irréalistes. Il faut alors ramener la solution à un niveau beaucoup plus modeste. Or, un tel geste met au jour les prétentions (fausses à notre avis) de cette solution élégante qui repose sur des exigences de fiabilité irréalistes. Nous nous interrogeons principalement sur le bien-fondé de l'utilisation de "fonctions en escalier" pour illustrer le degré de fiabilité "requis"; suivant ce genre de fonctions, toute variance inférieure à V^2 (valeur requise) prend une valeur constante et toute variance supérieure à V^2 prend une valeur nulle.

Une méthode très différente de celle décrite ci-dessus consiste à établir un genre de *moyenne minimisant la variance combinée (pondérée)* soit pour le coût fixe ou la taille d'échantillon fixe. Certes, si la variance ainsi obtenue est trop élevée (ou trop faible), on peut modifier les solutions en conséquence (c.-à-d. les relever ou les rabaisser). Nous préférons cette dernière méthode, qui est un compromis entre diverses répartitions qui existent chacune sous une forme optimale pour un seul usage (Yates 1981; Dalenius 1957). Cette méthode prévoit l'attribution de poids relatifs I_g à tous les paramètres statistiques énumérés, exercice qui peut paraître difficile à première vue (quoiqu'un décideur "inexpérimenté" pourra attribuer la même valeur à tous les paramètres). Toutefois, les deux autres solutions qui s'offrent à nous sont plus rigoureuses et devraient s'avérer encore plus difficiles à appliquer: la première consiste à définir le degré de fiabilité "requis" de tous les paramètres statistiques pour la première méthode, puis à attribuer arbitrairement des poids de même valeur à tous ces paramètres; la seconde méthode consiste à attribuer le poids unitaire à un paramètre statistique en particulier et des poids nuls à tous les autres paramètres.

En outre, il est facile de montrer que des compromis relatifs à la moyenne sont ordinaires, même réalisables et utiles parce que les répartitions ne sont pas influencées par les variations modérées de poids (comme c'est souvent le cas en statistique). Après tout, il est plus raisonnable de modifier l'importance relative d'une variable dans un rapport de 2 ou 5 par exemple que d'attribuer un poids de 1 à une variable et un poids nul à toutes les autres variables, cette dernière solution produisant des rapports de poids de valeur infinie.

Premièrement, définissons $\Sigma_i V_{gi}^2/n_i$ comme la variance possible d'un paramètre statistique g avec les résultats n_i de la répartition de la taille de l'échantillon pour la même composante de la variation. Ensuite, posons $1 + L_g(n) = (\Sigma_i V_{gi}^2/n_i)/V_g^2(\min) = \Sigma_i C_{gi}^2/n_i$ comme le taux d'accroissement (avec le résultat n_i de la répartition) de la variance du paramètre statistique g par rapport à la valeur minimum de cette variance; dans les deux cas, nous considérons la même valeur fixe Σn_i . Par conséquent, $L_g(n)$ représente la perte par rapport à la valeur minimum 1 ; à ce stade-ci, nous devons décider si les fonctions à minimiser seront les variances relatives C_{gi}^2/n_i ; il s'agit là d'une décision fondamentale. À notre avis, il n'y a pas de fonction plus appropriée que celles-là pour les besoins de l'équation (5.1) ci-dessus. Par exemple, nous préférons ces fonctions à V_{gi}^2 , qui repose sur des unités de mesure arbitraires, lesquelles sont effacées par $V_g^2(\min)$. Il arrivera toutefois à de rares occasions que $V_g^2(\min)$ soit nul ou très faible, auquel cas C_{gi}^2 pourrait être très élevée et instable; si cette situation devait se présenter, on attribuerait des valeurs arbitraires à C_{gi}^2 ou à L_g dans l'équation ci-dessus. Ces paramètres et ceux que nous verrons plus loin, y compris ceux du tableau 3, sont définis et analysés par Kish (1976).

Alors, étant donné les poids I_g servant à mesurer l'importance relative du paramètre statistique g pour n importe quelle série de résultats n_i de la répartition des tailles d'échantillons, nous avons:

$$1 + L(n) = \Sigma^g I_g (1 + L_g(n)) = \Sigma^g I_g \Sigma_i C_{gi}^2/n_i$$
$$= \Sigma_i \Sigma^g I_g C_{gi}^2/n_i = \Sigma_i Z_i^2/n_i$$

(1)

Après avoir modifié l'ordre des sommations, nous avons créé la nouvelle variable $Z_i^2 = \Sigma^g I_g C_{gi}^2$. Nous pouvons minimiser cette fonction afin d'obtenir des solutions de compromis pour

des erreurs d'échantillonnage pour les sous-classes et pour la comparaison des moyennes des sous-classes. Faisons en sorte que ces réponses soient plus explicites dans les plans de sondage à venir.

5. RÉPARTITION ENTRE LES DOMAINES

Cette source d'incompatibilité des plus importantes et des plus fréquentes revêt plusieurs aspects. Considérons tout d'abord la répartition de la taille globale d'un échantillon (ou la répartition des efforts ou des coûts) entre des domaines qui correspondent chacun à une partie de la population. Un exemple type est la répartition de la taille d'un échantillon entre les nombreuses (5, 10, 20 ou 50) provinces ou régions ou les nombreux états d'un pays; ces domaines renferment habituellement des populations N_d très différentes (dont le rapport peut varier de 1 à 100) même si les territoires eux peuvent être de même superficie ou presque. La question que l'on se pose souvent dans les circonstances est la suivante: les tailles d'échantillons (n_d) devraient-elles être à peu près les mêmes? Ou devraient-elles être proportionnelles à N_d avec des taux de sondage constants $f_d = f$? Des tailles identiques tendent à produire des erreurs de même valeur pour les moyennes (c.r.(\bar{y}_d)). En revanche, des taux de sondage constants tendent à produire l'erreur type la plus faible pour la moyenne globale $\bar{y}^w = \sum W_d \bar{y}_d$ parce qu'ils produisent des erreurs plus faibles pour les grands domaines. L'erreur type obtenue peut être moins élevée que la "normale", surtout si l'on tient compte des biais potentiels (figure 1), et peut ne pas justifier la taille élevée des échantillons et les coûts élevés. C'est ce que prétendent ceux qui préconisent la même taille d'échantillon (n_d) pour toutes les provinces. Cependant, on relève des erreurs d'échantillonnage plus élevées pour \bar{y}^w dans la plupart des sous-classes, notamment celles formées par le recouplement de variables telles l'âge, le sexe, la classe socio-économique, etc., et dont l'effet tend à être proportionnel à l'effectif total. Cela est le genre d'inconvénients qui découlent de la forte disparité des taux de sondage $f_d = n_d/N_d$ applicables aux provinces à cause de l'uniformité des tailles d'échantillons n_d .

Dans la Current Population Survey des États-Unis par exemple, on attribue un taux de sondage plus élevé aux États moins peuplés. La pondération qui s'ensuit a pour effet d'accroître la variance (pour un coût total fixe) des moyennes globales et des "classes de recoupement" comme celles formées des jeunes hommes et des jeunes femmes, et plus particulièrement des adolescents et adolescentes de race noire (touchés par des taux de chômage extrêmement élevés). Ce genre d'incompatibilité entre les besoins nationaux et régionaux (ou provinciaux) se retrouve dans tous les pays à cause des fortes variations de population d'une région à l'autre. Ainsi, le fait de vouloir accroître la qualité des données provinciales, pour un coût total fixe, est en contradiction avec le fait de vouloir accroître le degré d'exactitude des données nationales et des données recoupées. Pour réduire les risques de confusion, nous définissons les "domaines" comme des sous-ensembles de la population et les "sous-classes" comme des sous-ensembles de l'échantillon. Nous faisons ensuite la distinction entre les "domaines du plan" (et les sous-classes du plan), qui sont des sous-ensembles (par ex.: provinces et régions) contenus dans des strates définies par le plan de sondage, et les "classes de recoupement" (par ex.: âge, sexe, profession, revenu, etc.), qui s'étendent souvent presque aléatoirement sur plusieurs strates à la fois. Les effets du plan ne sont pas les mêmes pour ces deux genres de sous-classes (Kish 1961, 1980, 1987).

Par ailleurs, des difficultés peuvent surgir au point de vue de la répartition des variables et de la précision des variances $D^2 S_d^2$ pour des raisons autres que les différences de taille entre les "domaines"; cependant, il n'est pas approprié ici d'approfondir cet aspect de la question. Nous voulons plutôt attirer l'attention sur deux méthodes techniques qui permettent de résoudre simultanément tous les problèmes de répartition (la quatrième étape décrite à la fin de la Section 2). L'une de ces méthodes consiste à utiliser la programmation non linéaire itérative pour satisfaire simultanément à toutes les exigences de fiabilité relatives aux usages énoncés et ce, à un coût minimum. Cette solution élégante, qui convient à divers problèmes, fait appel à l'ordinateur et a fait l'objet de nombreux articles depuis 1963 (voir les comptes rendus et les bibliographies dans Bean et

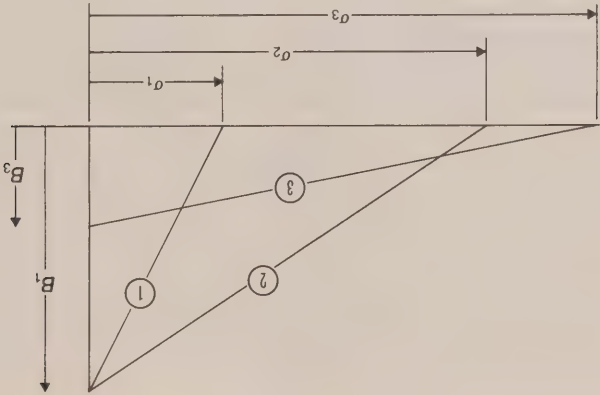


Figure 1. Erreurs variables (σ) et biais (B) dans l'erreur quadratique moyenne (EQM)

Les bases des triangles représentent les erreurs d'échantillonnage et d'autres erreurs variables (σ). Par exemple, σ_1 peut désigner l'erreur type de la moyenne \bar{y} de l'échantillon intégral, $ste(\bar{y}_1)$ peut désigner l'erreur type (plus élevée) d'une moyenne de sous-classe ($ste(\bar{y}_c)$) et σ_3 peut désigner l'erreur type de la différence entre deux moyennes de sous-classes ($ste(\bar{y}_c - \bar{y}_d)$). La hauteur des triangles représente les biais (B) et l'hypoténuse représente l'EQM = $\sqrt{(\sigma^2 + B^2)}$. (1) Pour l'échantillon intégral, le biais B_1 peut être élevé par rapport à l'erreur variable σ_1 ; par conséquent, le fait d'utiliser de plus grands échantillons ne réduirait pas l'EQM de beaucoup. (2) Par ailleurs, avec le même biais B_1 mais un échantillon moindre, tiré d'une sous-classe, le rapport change et l'erreur variable (σ_2) constitue désormais la majeure partie de l'erreur quadratique moyenne (EQM₂); du reste, EQM₂ n'est pas beaucoup supérieure à EQM₁ même s'il s'agit d'un échantillon beaucoup plus petit. (3) En ce qui concerne la différence de moyennes, le biais net B_3 peut être beaucoup moindre du rapport de biais B/σ tend à se produire non seulement dans le cas des différences de moyennes de sous-classes à l'intérieur d'un même échantillon, mais aussi dans le cas de différences de valeurs entre des enquêtes successives.

nous constatons que l'erreur type est environ trois fois plus élevée que dans le premier cas. Si nous comparons les moyennes de deux sous-classes de ce genre, l'erreur type (σ_3) est environ 1,4 fois plus élevée que dans le second cas. En revanche, nous pouvons constater que les hypothèses, qui représentent l'EQM, augmentent beaucoup moins fortement. Pour ce qui est de l'EQM₁, le biais B_1 est clairement supérieur à l'erreur type et cela peut être observé pour certaines variables dans de grands échantillons. Par contre, le fait de conserver le même biais ($B_2 = B_1$) pour la sous-classe a fait s'accroître l'EQM modérément et nous voyons que l'erreur type (σ_2) est maintenant supérieure au biais. La supériorité de l'erreur type est encore plus forte dans le cas de EQM₃, où σ_3 est supérieur à σ_2 mais où le biais (qui est censé être affecté du même signe qu'auparavant parce qu'il s'agit d'une tendance courante) est tombé à B_3 (comparaison des moyennes).

Des exemples de ces phénomènes se retrouvent partout et figurent à toutes fins dans le tableau 1. Nous choisissons l'exemple le plus courant — statistiques du chômage, où les biais observés peuvent éclipser entièrement les plus faibles valeurs des variations mesurables (par ex. 0,1 %). Toutefois, en ce qui a trait aux petites sous-classes (par ex. : adolescents de race noire), les erreurs d'échantillonnage pour les petites bases de sondage excèdent les biais. Dans le cas des comparaisons périodiques, les variations d'échantillonnage sont encore plus prononcées. Les rapports que nous venons de décrire et qui mettent en relation les biais et les erreurs variables ne sont pas imposés par des considérations logiques; ils relèvent plutôt de considérations empiriques et courantes. Le fait de négliger ces rapports élémentaires créé beaucoup de confusion quant à la nécessité d'avoir des enquêtes par sondage d'une grande précision (c.-à-d., erreurs d'échantillonnage faibles(σ)). Par la figure 1, nous voulons donner des réponses pratiques à des questions courantes comme celles qui suivent: pourquoi dépenser de l'argent pour de grands échantillons et des méthodes d'échantillonnage rigoureuses devant des biais de mesure élevés? Pourquoi se donner la peine de calculer les erreurs d'échantillonnage si les biais de réponse constituent la majeure partie de l'erreur totale? Les réponses à ces questions se trouvent dans la supériorité

Tableau 2

Dix sources d'incompatibilité (a-j)

a.(4) Des tailles m_g ou des taux f_g sont nécessaires pour les usages g	
$V_g^2 = S_g^2 D_g^2 / m_g$ et $m_g = S_g^2 D_g^2 / V_g^2$ ou $f_g = S_g^2 D_g^2 / V_g^2 P_g$	
où m_g désigne la taille des sous-classes et $f_g = n_g / N = m_g / P_g$ désigne le taux de sondage	
b. (4) Rapport entre les biais et les erreurs d'échantillonnage dans $EQM = \sqrt{(\sigma^2 + B^2)}$	
- Le rapport de biais B/σ diminue lorsque σ s'accroît pour les sous-classes	
- À des fins de comparaison, B/σ tend à diminuer lorsque B diminue ou σ augmente	
c. (5) Répartition de m_g entre les domaines	
$m_i = \sum_g m_g$	
d. (6) Répartition de m_{gh} entre les strates h	
$m_g = \sum_h m_{gh}$	
e. (6) Choix de variables pour la stratification	
Stratification à plusieurs variables	
f. (7) Tailles de grappe optimales	
$D_g^2 = [1 + \rho_g(b_g - 1)]b_g = P_g n_g / a$ pour les classes de recoupement	
g. (7) Mesures pour les tailles de grappe	
h. (7) Conservation des unités d'échantillonnage (UPÉ) pour de nouveaux sujets, de nouvelles mesures et de nouvelles strates et pour des sujets variés.	
i. (8) Évolution du plan de sondage	
Quel est le degré de chevauchement? Doit-il y avoir des panels? Variation vs agrégation.	
j. (9) Calcul et présentation des erreurs d'échantillonnage.	
Les chiffres (4) à (9) renvoient aux sections où sont traités ces problèmes.	

Pour la comparaison de sous-classes, on observe une augmentation encore plus forte des variances: $m_g = (m_a^2 + m_b^2)^{-1} = n(P_a^{-1} + P_b^{-1})^{-1}$, où P_a et P_b désignent des proportions de l'échantillon n (en supposant que $S_a^2 = S_b^2$). Si nous comparons par exemple deux moyennes de sous-classes ($0.01n$ et $0.10n$), nous avons la "taille effective" $m_g = n(0.01^{-1} + 0.10^{-1})^{-1} = n/1.10$. Pour ce qui a trait à d'autres paramètres statistiques comme les médianes et les coefficients de régression, il serait trop complexe de définir des tailles d'échantillon "requises". Cela dépasse le cadre de notre étude mais il n'est pas exclu que nous présentions certains chiffres.

L'analyse des paramètres statistiques des sous-classes prend une toute autre dimension si nous incluons le biais B^2 dans l'erreur quadratique moyenne ($EQM = RMSE = \sqrt{(\sigma^2 + B^2)}$). La figure 1 sert à illustrer une tendance courante dans l'exactitude des données d'enquête bien qu'il puisse y avoir des écarts appréciables entre les rapports du biais à l'erreur d'échantillonnage; nous recommandons fortement de lire la légende qui accompagne la figure. Il arrive souvent que le biais potentiel B^1 soit supérieur à l'erreur d'échantillonnage mesurable σ_1 pour tout l'échantillon. Cependant, si nous prenons une sous-classe qui représente environ le dixième de l'échantillon,

n'est probablement pas exhaustive et le lecteur pourrait bien lui ajouter des éléments. Mieux encore, il pourrait définir à l'intérieur de chacun des points du tableau des problèmes et des solutions qui ne sont pas traités dans le présent document. Il serait souhaitable de mettre les dix sources d'incompatibilité en relation logique avec les vingt usages répartis en six niveaux; par exemple, on a dix noeuds niveau/source d'incompatibilité. Malheureusement, les sources d'incompatibilité dénotent une dimension perpendiculaire par rapport à l'usage et les soixante (10×6) cellules (ou la plupart d'entre elles) ont un contenu significatif.

Heureusement, les dix sources d'incompatibilité ne sont pas toutes présentes dans chaque plan de sondage. Nous estimons qu'il faudrait toujours considérer, du moins de façon informelle, la possibilité d'un problème au point de vue des tailles d'échantillon m_g et du rapport entre les biais et les erreurs d'échantillonnage parce qu'il s'agit là des problèmes les plus fréquents. Il faudrait aussi se pencher, au moins brièvement, sur la répartition de la taille de l'échantillon entre les domaines et la répartition entre les strates; de fait, il faudrait s'arrêter encore plus souvent sur ces points. Le calcul des erreurs d'échantillonnage (i) devrait aussi être examiné dans la plupart des enquêtes. En revanche, pour ce qui a trait aux enquêtes uniques, on ne s'intéressera pas aux problèmes qui touchent la permanence du plan de sondage. Par contre, dans le cas d'une enquête permanente fondée sur une base de sondage permanente, les décisions concernant e), f), g) et h) (stratification, taille des grappes et mesures) peuvent avoir été prises il y a longtemps pour un plan de sondage fixe. Malgré cela, les tailles de grappe (f) utilisées dans les étapes intermédiaires (ilots et segments) peuvent faire l'objet de modifications opérationnelles souples.

Il est de plus rassurant de savoir que des compromis fondés sur des méthodes statistiques peuvent produire des résultats tout à fait acceptables et ce, pour plusieurs raisons (Section 5 à 8). La première est qu'un faible écart entre la répartition observée et la répartition optimale n'entraîne qu'un très faible accroissement de la variance. Les courbes d'efficacité tendent à être horizontales sur de larges intervalles centrés sur les points optimaux; ainsi, une grande précision pour des plans de sondage distincts, par ailleurs impossible à obtenir, n'est pas nécessaire. La seconde raison est qu'un écart appréciable entre la répartition observée et la répartition optimale peut entraîner un accroissement notable de la variance. Ainsi, le fait d'ignorer les usages importants peut entraîner des pertes d'efficacité appréciables dans leur cas, ce qui nous oblige à élaborer des plans de compromis pour ces usages. La troisième raison est que les plans de compromis, en conformité avec les méthodes statistiques, peuvent réduire sensiblement les pertes élevées que pourrait engendrer une optimisation des répartitions pour d'autres usages, et produire des variantes à peine plus élevées que celles découlant des plans optimaux pour chaque usage (Section 5).

4. TAILLE DES ÉCHANTILLONS ET RAPPORTS DE BIAIS (B/σ)

Ces deux sources d'incompatibilité (a et b dans le tableau 2) passent probablement pour les plus importantes de toutes à cause des conséquences qu'elles peuvent avoir. Si nous les considérons ensemble ici, c'est uniquement parce qu'elles peuvent être liées étroitement l'une à l'autre par les effets des sous-classes. Commençons par le cas le plus connu (échantillonnage aléatoire simple avec remise), où $m = S^2/V^2$ représente la taille d'échantillon nécessaire pour obtenir un degré de précision "requis", V^2 pour une moyenne d'échantillon y avec variance S^2 des unités élémentaires. Or, S^2 dépend largement des variables et des domaines, qui varient en fonction de g pour la moyenne \bar{y}_g , et la valeur "requis" V^2 peut varier d'avantage. Nous considérons aussi les effets de plan D_g^2 , qui varient également, et par voie de conséquence, $m_g = S_g^2 D_g^2 / V^2$ représente la taille d'échantillon nécessaire pour la moyenne de la variable g. En ce qui concerne la moyenne \bar{y}_g d'un domaine g qui ne comprend qu'une proportion P_g de la population, la taille d'échantillon *globale* requise pour le domaine devient $n_g = m_g / P_g$ et il est plus pratique d'exprimer la fraction de sondage requise sous la forme $f_g = n_g / N = S_g^2 D_g^2 / V^2 P_g N$. On peut omettre le facteur (1-f) ou l'inclure dans D_g^2 . P_g devient faible et déterminante s'il est nécessaire d'avoir un degré de précision élevé pour les petites sous-classes.

Tableau 1
Hiérarchie d'usages

1. Statistiques diverses établies à partir des mêmes variables
 - Totaux, moyennes, médianes et quantiles; distributions
 - Statistique analytique: régressions, analyses de catégories
 - Aspects chronologiques; statique, macro-variation, micro-variation, cumulatif
2. Populations et domaines divers (sous-classes)
 - Classes propres et classes de recoupement
 - Comparaison de sous-classes
3. Variables multiples sur le même sujet
 - Différentes mesures d'une même variable (par ex.: revenu, chômage)
 - Périodes diverses — par jour, par semaine, par mois, par année
 - Plusieurs aspects d'un même sujet: revenu, épargne, patrimoine
4. Enquêtes portant sur plusieurs sujets
 - Plusieurs sujets traités sur le même questionnaire, durant la même interview ou au cours de la même enquête
 - Enquêtes-santé portant sur de nombreuses maladies
 - Etudes de marché pour de nombreux clients et portant sur de nombreux produits
 - Enquêtes agricoles portant sur de nombreuses cultures
 - Enquêtes sociales à usages multiples
5. Opérations d'enquête intégrées (enquêtes permanentes)
 - La NSS en Inde, la CPS aux E.-U., le NHSCP des N.-U.
 - Enquêtes distinctes confiées à des divisions et à des bureaux régionaux différents
 - Source commune des enquêtes
 - Diversité des méthodes, des coûts, des opérations, des répartitions, des répondants
6. Bases de sondage principales
 - Plusieurs échantillons tirés d'une base de sondage ou d'une série de listages
 - Établissements, organismes distincts
 - Effectifs régionaux distincts? Mêmes UPE?

liste complète de ces termes statistiques. Deuxièmement, il faut connaître les estimations des *facteurs de variance et des facteurs de coût* pour chaque usage. Troisièmement, certaines méthodes exigent que l'on établisse des valeurs au degré de précision "requis" pour tous les usages (Section 5). Quatrièmement, les valeurs et les estimations précitées doivent être incluses dans une formule mathématique qui aura pour solution un plan "optimal" unique qui sera prêt à être utilisée. Les ordinateurs ont simplifié les calculs nécessaires à une telle solution mais l'analyse conceptuelle et théorique est toujours aussi exigeante (Section 5).

Les exigences qui se rattachent à l'étude des plans à usages multiples nous aident à comprendre pourquoi les manuels font peu souvent mention de ce genre de plans. Il convient toutefois de signaler les ouvrages qui figurent dans la bibliographie ci-jointe de même que les notes et la bibliographie qui paraissent dans Rodriguez-Vera (1982), Cochran (1977) et Chattejee (1967). Par ailleurs, les descriptions d'enquêtes présentent souvent un paramètre statistique (par ex., la moyenne) d'une variable principale comme le seul objet (principal) de l'enquête. Dans le contexte des plans à usages multiples, cela revient à enlever toute importance aux autres buts de l'enquête. On peut atténuer l'invasivité de cette affirmation en supposant que les autres objets principaux donneraient lieu à des répartitions semblables; il n'en reste pas moins que cette affirmation devrait être étayée à l'aide des calculs prévus dans les quatre étapes exposées précédemment.

3. APERÇU DE DIX SOURCES D'INCOMPATIBILITÉ

Il serait utile d'examiner brièvement les dix sources d'incompatibilité énumérées dans le tableau 2 avant d'aborder des problèmes spécifiques et de proposer des solutions. La liste du tableau 2

et à les sensibiliser à l'importance et à la réalité de ces plans. En second lieu, nous voulons poser les jalons d'une recherche intégrée sur les nombreux problèmes que soulèvent les plans de sondage à usages multiples. Souhaitons que les imperfections de nos méthodes incitent d'autres personnes à poursuivre la recherche et à élaborer de nouvelles méthodes.

2. UNE HIÉRARCHIE D'USAGES

Au départ, il nous faut clarifier le sens de l'expression "à usages multiples" car celle-ci renvoie à un trop grand nombre de concepts dans la documentation statistique. À cette fin, nous avons réparti une vingtaine d'usages en six niveaux hiérarchiques (tableau 1). Les niveaux 3 (variables multiples) et 4 (enquêtes portant sur plusieurs sujets à la fois) sont ceux qui font le plus souvent l'objet d'analyses et parfois, ils sont considérés séparément pour les mêmes variables ou des variables étroitement liées (Murthy 1967). Chacun des six niveaux est subdivisé selon les formes d'usage, celles-ci étant analysées plus en détail dans d'autres ouvrages (Nations-Unies 1980; Lahiri 1963); les listes du tableau 1 ne sont pas exhaustives.

Les opérations d'enquête intégrées (niveau 5) ont un lien avec les enquêtes portant sur plusieurs sujets mais doivent être distinguées de celles-ci parce qu'elles se rattachent à des organismes et à des établissements qui réalisent de nombreuses enquêtes dans divers domaines sur de plus longues périodes (Nations-Unies 1980; Foreman 1983). Le cinquième niveau était désigné comme les "opérations d'enquêtes permanentes" jusqu'à ce qu'on reconnaisse que la plupart des grandes enquêtes par sondage étaient réalisées par des organismes d'enquête permanents comme le U.S. Census Bureau, Statistique Canada ou le Survey Research Center. Cette permanence présente des avantages notables au point de vue des coûts et de la qualité, sans compter les effets restrictifs sur les plans de sondage (Kish 1965).

Les bases de sondage principales ou les échantillons principaux (niveau 6) représentent une forme plus poussée et plus spécialisée de plans de sondage à usages multiples. Cela peut tout simplement signifier que l'on utilise les mêmes cartes ou listes d'îlots ou les mêmes segments de secteur pour des enquêtes différentes ou que l'on recourt à l'"échantillon principal de l'agriculture" (King et Jessen 1945), méthode selon laquelle les régions rurales de tous les comtés des E.-U. sont divisées en segments comptant environ quatre fermes chacun; il peut aussi s'agir de l'entreprise qui vend la liste courante des logements pour la formation de la plupart des échantillons utilisés en Allemagne de l'Ouest. Ces exemples très variés ont toutefois quelque chose en commun: dans les trois cas, le partage des "fraîs de mise en marche" (conception du plan, stratification, listage, etc.) pour la construction de bases de sondage permet de réaliser des économies.

Les ouvrages portant sur l'échantillonnage à usages multiples ont toujours passé sous silence les paramètres statistiques fondés sur une seule variable et des domaines variés (niveaux 1 et 2) même si ces catégories sont les plus courantes; toutefois, comme nous le verrons plus loin, ces catégories peuvent avoir les conséquences les plus notables et soulever les problèmes les plus sérieux. L'effet du plan sur des paramètres statistiques comme les médianes, les quantiles et les coefficients de régression peut être très différent de l'effet du même plan sur les moyennes et les agrégats (Kish 1961; Kish 1965; Kish et Frankel 1974). De plus, la formation d'échantillons périodiques soulève de nouvelles questions (Section 8). Toutefois, l'effet du plan se fait sentir surtout sur les moyennes des petites "sous-classes" (c'est-à-dire, aussi petites que 0,10 ou 0,01) de l'échantillon intégral, ces sous-classes correspondant à des "domaines" similaires dans la population (Section 5). Chacun des niveaux du tableau 1 présente des aspects différents des plans de sondage et peut être approfondi avec profit afin d'en préciser la signification et d'en tirer des exemples plus précis (dont certains figurent déjà dans le tableau 1).

Les lacunes qui ont valu aux plans de sondage à usages multiples d'être omis sont nombreuses. Premièrement, il est nécessaire de formuler les différents usages *en termes statistiques explicites* de sorte qu'ils puissent être utilisés dans les formules qui permettent de les comparer et celles qui servent à établir des compromis mais voilà, la principale difficulté est peut-être d'obtenir une

Plans de sondage à usages multiples¹

LESLIE KISH²

RÉSUMÉ

La plupart des enquêtes ont de nombreux usages et nous proposons dans cet article une hiérarchie de ces usages en six niveaux. Toutefois, la plupart des théories et des ouvrages statistiques mettent l'accent sur les enquêtes à usage unique pour éviter la complexité des plans de sondage à usages multiples et les problèmes que ces plans soulèvent. Nous exposons tout d'abord dix sources d'incompatibilité entre les usages de ces plans, puis nous posons des problèmes et donnons des solutions pour chaque cas. Heureusement, des compromis et des solutions communes sont possibles puisque la plupart des optimums sont très peu prononcés et que la plupart des "exigences" relatives à la précision sont en réalité très souples. Il est préférable de parler d'usages multiples et de composer avec eux que de se limiter à quelques usages choisis arbitrairement; en outre, la venue de l'informatique a facilité l'application des plans de sondage à usages multiples.

MOTS CLÉS: Répartition entre les domaines; erreurs quadratiques moyennes; répartition à usages multiples; plan à usages multiples; répartition optimale; échantillons périodiques; taille de l'échantillon.

1. INTRODUCTION

Dans la plupart des cas, la planification des enquêtes révèle plusieurs usages, puis beaucoup d'autres usages apparaissent au moment de l'analyse des données, et d'autres encore au moment de leur interprétation et de leur utilisation. Toutefois, la nature polyvalente de la plupart des enquêtes tend à être dissimulée par des analyses de plans d'enquête trop simplifiées, qui ne tiennent compte que d'une variable. Cette lacune ressort particulièrement dans le cas des enquêtes par sondage, auxquelles nous nous intéresserons dans le présent article, mais elle est observable aussi dans le cas d'autres plans statistiques comme les études expérimentales et les études d'évaluation. En pratique, les enquêtes visent habituellement plusieurs objectifs. Pourquoi alors la théorie de l'échantillonnage ne traite-t-elle pas les plans de sondage à usages multiples? Parce qu'une théorie des plans à usages multiples serait trop complexe et que la théorie de l'échantillonnage est déjà assez complexe comme cela; nous relèverons plus loin des exceptions particulières. Par surcroît, les descriptions de plans de sondage que nous lisons couramment tendent à refléter le caractère "prestigieux" de la théorie de l'échantillonnage à usage et à variable unique au lieu de représenter fidèlement les nombreux compromis auxquels il faut arriver dans la réalité. De nombreux plans de sondage courants (notamment, la méthode d'échantillonnage avec probabilités égales) servent probablement à de nombreux usages de façon efficace; selon nous, la conception *explicite* de plans de sondage à usages multiples semble être une chose rare mais combien nécessaire. La *polyvalence* des échantillons d'enquête revêt plusieurs aspects qui sont hiérarchisés en six *niveaux* dans la section suivante. Nous définissons ensuite *dix sources d'incompatibilité* entre des objectifs. Ces sources d'incompatibilité sont traitées séparément dans les sections 4 à 9, où nous présentons des façons de résoudre l'incompatibilité. Certaines des solutions proposées puisent dans des éléments bien connus de la théorie des sondages tandis que d'autres sont plus inédites, donc moins bien élaborées et documentées.

Par cet article, nous visons tout particulièrement à fournir aux utilisateurs de la documentation utile sur les méthodes de conception et d'application des plans de sondage à usages multiples

¹ Communication-thème présentée au Symposium international sur la statistique de Taipei, Taiwan, août 1986, et à un colloque de Statistique Canada, 7 octobre 1987.

² Leslie Kish, Institute for Social Research, University of Michigan, Ann Arbor, MI 48104 E.-U.

i) la composante c.a.f. des importations doit faire l'objet d'une mesure systématique. Sinon, il ne serait pas possible de comparer les exportations et les importations en général. Les renseignements deviennent disponibles au moment où l'importation est déclarée aux Douanes. Les questions telles que la fréquence et le niveau de détail dépendront des ressources et de l'urgence à améliorer les connaissances des utilisateurs;

ii) il faudrait entreprendre une étude des retards entre les exportations et les importations par catégorie de marchandises et par pays d'origine. Pour que cette étude soit rentable, il faudra probablement faire appel à la collaboration des pays partenaires, mais si cette étude n'est pas prévue dans un avenir rapproché, on pourrait utiliser comme remplacement acceptable les factures commerciales;

iii) à partir de ces deux éléments, il faudrait utiliser une méthode officielle d'estimation des importations correspondantes sur la base des exportations, et totaliser l'erreur d'estimation pour étude ultérieure. Si l'erreur d'estimation n'a pas d'autocorrélation appréciable, les erreurs de codage et connexes pourraient expliquer la différence entre l'importation enregistrée et son estimation statistique. Si, par contre, le terme d'erreur ne répond pas à ces critères, il faudrait en faire l'objet d'une étude ultérieure en collaboration avec le pays partenaire;

iv) il faudrait vérifier les excédents ou les déficits évidents par rapport aux pays qui ont probablement un rôle d'intermédiaire commercial ou d'entrepôt. Ainsi, un excédent d'exportation avec les Pays-Bas pour les États-Unis pourrait faire l'objet d'une vérification par rapport aux déficits correspondants avec des pays tels que la République fédérale d'Allemagne ou la France. Les méthodes économétriques peuvent servir à séparer l'effet global des services d'entrepôt, même si ces derniers vont servir plus probablement aux marchandises volumineuses et entroposables, des effets de courte durée tels qu'une erreur de codage.

v) pour les marchandises qui sont des valeurs systématiquement des valeurs aberrantes, une fois toutes les corrections faites, parce qu'elles persistent dans le temps ou parce qu'elles touchent tous les pays, il faudrait se servir du Système harmonisé et obtenir l'aide du Conseil de coopération douanière pour l'interprétation de ses notes explicatives.

Il est évident que le lancement d'un tel programme nécessite des préparatifs, les approbations nécessaires et des ressources. Il ne peut être exécuté immédiatement, et la plupart des pays ne vont pas le parrainer tout de suite. Mais les propositions ne devraient pas être laissées de côté, comme cela a été le cas il y a quelque treize ou quatorze ans pour des propositions semblables. On accorde beaucoup trop d'attention aux statistiques du commerce pour prendre le risque de retarder leur amélioration. La comparaison avec les données correspondantes révèle qu'on ne pourra y accorder plus d'attention que si elles sont sensiblement améliorées, ou si leurs analystes deviennent plus conscients des limites des données qui servent à tester leurs hypothèses.

BIBLIOGRAPHIE

- ALLEN, R.G.D., et ELY, J. EDWARD (1953). *International Trade Statistics*, New York: John Wiley & Sons.
- COATS, R.H. (1926). *Canadian Trade Statistics. Journal of the Canadian Bankers' Association*.
- RYTEN, J. (1983, 1984, 1986). Reports on the Reorganization of Bolivian Foreign Trade Statistics (En espagnol). United Nations Development Programme, New York.
- UNITED NATIONS (1982). *International Trade Statistics, Concepts and Definitions. Statistical Papers Series M, 52*, New York.
- UNITED NATIONS (1986). *Standard International Trade Classification. Statistical Papers Series M, 34*, New York.
- UNITED NATIONS STATISTICAL COMMISSION (1974). *International Trade Reconciliation Study*. Rapport du Secrétaire Général. Dix-huitième session, Genève.

Pour trente rapports x/m (en comptant les trois rapports pour la CFE dans son ensemble), on relève neuf cas (chiffres suivis d'un astérisque dans le tableau) pour lesquels les prédictions ne se vérifient pas. En supprimant les deux chiffres pour la Grèce, parce que les flux commerciaux correspondent très difficilement à l'étranger, pour accroître le niveau de détail de la classification par marchandise et augmenter le nombre de ventilations grâce à des variables de classification supplémentaires. Même lorsqu'on les regroupe dans le temps, les transactions de ces cellules détaillées ne correspondent guère avec leurs homologues. Comme on ne peut affirmer que les deux rapports d'une comparaison bilatérale sont simultanément corrects, il est possible que les deux contiennent une erreur appréciable.

4. SIGNALER LES ERREURS AUX UTILISATEURS

Il y a deux questions distinctes. L'une est d'informer les utilisateurs que, contrairement à une opinion répandue, les chiffres du commerce extérieur, et en particulier les chiffres détaillés, peuvent être biaisés. L'autre est de mettre sur pied un programme d'amélioration de la qualité des données du commerce extérieur qui bénéficierait du fait que des mesures correspondantes de la même transaction existent. Voici quelques propositions pour lancer un tel programme. L'analyse contenue dans cet article prévoit qu'au-delà du niveau à deux chiffres de la classification par marchandise par pays, même annuellement, il n'est pas possible d'accorder une confiance complète aux niveaux ni aux variations d'une année à l'autre. Les utilisateurs auront probablement une réaction négative à cette conclusion, puisqu'ils ont déjà des raisons de mettre en doute le champ des agrégats dans le cas des exportations. Les résultats du programme de rapprochement de la statistique des Etats-Unis et du Canada ne doivent pas être considérés comme étant limités à ces deux pays. D'autres pays ont à faire face à la même catégorie de problèmes à des degrés divers. La révélation que, en plus de ces lacunes, les données par marchandise au-delà d'un certain niveau ne peuvent être utilisées qu'avec beaucoup de prudence, pourrait se traduire par un changement fondamental de la perception des utilisateurs de la statistique du commerce extérieur.

Mais, si cette mesure n'est pas adoptée, quel qu'en soit le degré d'impopularité, on maintient une croyance qui n'est pas entièrement justifiée par les faits. Les chiffres détaillés du commerce de marchandises servent à divers usages, dont le principal est celui des tarifs. Les discussions sur ces questions utilisent beaucoup les chiffres détaillés, rarement les différences entre les données nationales et correspondantes et aussi rarement la statistique de la consommation intérieure comme mesure de vérification de l'ordre de grandeur que semblent indiquer les données douanières. De plus, dans une autre utilisation de données détaillées des marchandises, on tire des opinions sur des politiques industrielles et régionales, et des mesures peuvent être prises à partir d'éléments de preuve qui, selon cette analyse, ne seraient pas solides. Il est évident qu'il appartient aux organismes statistiques de communiquer aux utilisateurs les lacunes relevées des données afin d'empêcher la généralisation d'une mauvaise utilisation.

5. UN PROGRAMME D'AMÉLIORATION DE LA STATISTIQUE DU COMMERCE EXTERIEUR

En plus de fournir aux utilisateurs plus de renseignements concrets sur les erreurs dans la statistique du commerce extérieur, il faudrait mettre sur pied un ou plusieurs programmes destinés à améliorer la qualité de ces statistiques dans le temps. Voici des mesures qui auraient du probablement être prises il y a quelque temps :

Tableau 6
Comparaison des statistiques du commerce extérieur correspondantes
pour deux années choisies

	1979	1983
Nombre de dossiers avec $x > m$ en pourcentage de tous les dossiers	35	32
Valeur des exportations avec $x > m$ en pourcentage du total des exportations	41	42
Rapport x/m pour $x > m$	1.18	1.15
Rapport x/m pour $x < m$.87	.85

Tableau 7
Rapports X/M en 1985
de trois pays déclarants sélectionnés à neuf partenaires commerciaux

A	De	Canada	Etats-Unis	Japon
C.E.E.		.84	.92	.94
Pays-Bas		1.93	1.34	1.33
Belgique - Luxembourg		1.47	1.51*	1.26
Danemark		1.20*	.74	1.05*
France		.70	.74*	.69*
République fédérale d'Allemagne		.69	.81	.98
Irlande		.55	.78	.72*
Italie		.75	.84	.75*
Royaume-Uni		.74	.86	.89
Grèce		1.00	1.23*	.89*

Le tableau 6 présente les variations entre deux années choisies d'un certain nombre d'indicateurs reliés à des cas où les exportations ont dépassé les importations correspondantes. Même si sur une période de quatre ans il y a eu certains changements dans le nombre de dossiers pour lesquels les exportations ont dépassé les importations ainsi que la valeur unitaire, les changements en question sont mineurs. Assez curieusement, les cas de x/m représentent plus de 40% de la valeur totale du commerce, et comme ce chiffre a augmenté légèrement, le nombre de dossiers qui l'explique a baissé de 10%.

Au tableau 7, on soumet à l'épreuve des faits un certain nombre de prédictions *a priori*. On étudie trois exportateurs déclarants, le Canada, les Etats-Unis et le Japon, et neuf partenaires commerciaux déclarants, c'est-à-dire les pays membres de la CEE autres que l'Espagne et le Portugal. Les tableaux donnent la liste des rapports simples x/m de 1985 pour le commerce pays à pays. Toutes choses étant égales par ailleurs, les prédictions suivantes semblent plausibles:

- plus la part de la fabrication d'un flux commercial est élevée, plus le rapport x/m est élevé, ce qui revient à dire que plus le rapport c.a.f./valeur totale est plus petit, et plus la valeur ajoutée incorporée dans une marchandise est élevée. Pour cette raison, le classement par ordre *croissant* des rapports devrait être Canada, Etats-Unis, Japon;
- dans le cas du commerce avec les pays d'entreposage, soit les Pays-Bas, et dans une moindre mesure, l'Union belgo-luxembourgeoise, le mauvais codage du pays par l'exportateur ne devrait toucher principalement que les livraisons en vrac. Pour cette raison, le rapport x/m par ordre décroissant devrait être Canada, Etats-Unis, Japon; et
- des rapports x/m supérieurs à un ne devraient s'observer que pour les pays d'entreposage.

On obtient les chiffres du tableau en prenant l'indice qui mesure la variation de chaque section (1 chiffre) du rapport simple X/M de 1978 à 1985, c'est-à-dire $(x/m)_{1985}$ divisé par $(x/m)_{1978}$, et en le divisant par un indice correspondant dans lequel le rapport normalisé (x/m) pour 1985 a été utilisé et où les rapports de divisions (2 chiffres) ont été agrégés à partir de leurs parts de 1978 dans leur division correspondante. Algébriquement, le rapport obtenu R_i est:

$$R_i = 100 \cdot M_{178} \cdot \frac{X_{i85}}{M_{i85}} \div \sum_{n_i}^{j=0} \frac{m_{ij85}}{x_{ij85}} \cdot m_{ij78}.$$

Un chiffre de 104, par exemple, signifie qu'une augmentation de quatre pour cent de la valeur courante des exportations par rapport aux importations correspondantes s'explique par des raisons autres que l'effet des changements de la composition des marchandises sur la composante c.a.f.

3.2 Analyses utilisant la base de données du commerce mondial complète

Possibilités de mauvaise classification par pays et marchandise:

Les tableaux 5 et 6 tirent de la base de données du commerce mondial complète présentent le nombre de cas de mauvaise classification possibles par pays et marchandise. Le tableau 5 contient le nombre de cas pour 1983 pour lesquels il y a des échanges bilatéraux d'une marchandise selon l'un des pays déclarants d'un couple partenaire commercial, mais non selon l'autre. Ce nombre de cas est exprimé pour chaque niveau de détail de la CTCl comme une proportion de tous les cas. Le tableau 5A présente l'impact sur la valeur, là encore pour chaque niveau de la CTCl. En plus de fournir une mesure sommaire de la taille des erreurs, les tableaux donnent aussi une idée de la rapidité avec laquelle le nombre de situations anormales augmente en fonction du niveau de détail de la classification.

Tableau 5

Comparaison des statistiques du commerce extérieur correspondantes en 1983 - Nombre de dossiers¹

Niveau de détail de la CTCl			
Pourcentage ne rapportant pas d'exportations		Pourcentage ne rapportant pas d'importations	
Pourcentage		Pourcentage	
du total		du total	
0 (combiné)	11	4	15
1 chiffre	14	7	21
2 chiffres	16	10	26
3 chiffres	19	13	32

Pourcentage du nombre de dossiers de couples partenaires commerciaux où un membre déclare 0 exportations/importations et les autres membres déclarent un autre chiffre

Tableau 5A

Comparaison des statistiques du commerce extérieur correspondantes en 1983 - Valeurs des dossiers¹

Niveau de détail de la CTCl			
Pourcentage ne rapportant pas d'exportations		Pourcentage ne rapportant pas d'importations	
Pourcentage		Pourcentage	
du total		du total	
0 (combiné)	.1	-	.1
1 chiffre	.3	.1	.4
2 chiffres	.6	.4	1.0
3 chiffres	1.1	.9	2.0

¹ Pourcentage de la valeur des dossiers des couples partenaires commerciaux, dont un membre déclare 0 exportations/importations et l'autre un commerce non nul

Le symbole (-) indique une valeur insignifiante.

Tableau 4

Variations des rapports simples x/m entre 1978 et 1985 comparées aux rapports normalisés x/m avec des parts de division CTCI constantes

Exportations de ...

à ...

A.N.

CÉE

CÉE

Japon

A.N.

Japon

Sections CTCI					
0	Aliments	107	102	100	98
1	Boissons et tabacs	99	100	99	99
2	Matières brutes	100	100	96	100
3	Combustibles minéraux	102	117	304	103
4	Huiles animales et végétales	99	98	107	100
5	Produits chimiques	102	101	101	100
6	Produits manufacturés	99	101	99	98
7	Machines et transport	96	91	95	92
8	Articles manufacturés divers	100	100	100	97
9	Transactions diverses	150	176	163	86

en supposant que les proportions des importations par section au total des importations pour chaque flux de marchandises demeurent constantes depuis 1978. Ces rapports normalisés sont une approximation d'une estimation qui élimine l'impact des variations dans l'ensemble du c.a.f. de la variation du rapport dans le temps. Toute différence entre les deux rapports de 1985 doit donc être attribuée à d'autres facteurs.

L'évolution des rapports dans le temps en raison de l'augmentation de la part des produits hautement manufacturés dans certaines exportations de flux est assez prévisible. Ainsi, les exportations de la CEE vers l'Amérique du Nord et le Japon, les exportations du Japon vers la CEE et l'Amérique du Nord devraient inclure proportionnellement davantage d'articles manufacturés. Le rapport qui traduit les variations de composition est donc supérieur au rapport normalisé, car l'importance relative du c.a.f. diminue à mesure que la valeur d'une unité de poids ou de volume augmente.

Mais il y a des exceptions *a priori* à cette certitude dans le tableau. Ainsi, les exportations de l'Amérique du Nord vers le Japon se caractérisent par un très grand écart entre le rapport simple et le rapport normalisé, même si la part des articles manufacturés a augmenté relativement moins. Le tableau 4 contient une ventilation par section de la CTCI pour les rapports correspondants aux flux commerciaux entre chacun des six couples de blocs commerciaux retracés dans la mini-base de données. Les chiffres indiqués sont des rapports de l'indice simple au niveau de la section à l'indice calculé à partir de la part des importations au niveau de la division. Ils indiquent l'apport de la variation des rapports représenté par les variations de la composition de marchandises. Ce ne sont rien d'autre que des indicateurs, en partie parce qu'ils diminuent d'un niveau dans la classification des marchandises. On ne peut facilement déceler de tendances, il y a en gros autant de cas de dépassement que de rabage.

Pour les flux importants, tels que ceux de l'Amérique du Nord vers la CEE ou de la CEE vers le Japon, la composition par marchandise est relativement stable, et il n'y a guère de différence entre les rapports à pondération de base et à pondération courante. De plus, ces rapports ne fluctuent guère au cours de la période étudiée. D'autres flux sont plus sensibles à la composition par marchandises, ce qui semble indiquer qu'à des niveaux plus bas de la classification les différences f.o.b./c.a.f. expliquent une petite partie de la variation des rapports x/m dans le temps.

Variation des rapports X/M entre 1978 et 1985 et comparaisons avec les rapports normalisés X/M , les parts des sections CTCl étant supposées constantes¹

Amérique du Nord		CÉE		Japon	
Rapport simple	Rapport normalisé	Rapport simple	Rapport normalisé	Rapport simple	Rapport normalisé
1978	1985	1978	1985	1978	1985
.96	.92	.96	.91	.95	.98
.92	.91	.90	.94	.86	.89

$$\text{Rapport simple} = \frac{1}{M_{78}} \sum_n^{i=0} \frac{m_{78}^i}{X_{78}^i} \cdot m_{i78},$$

qui regroupe, entre autres, tous les types de matériel de transport. La différence des taux de croissance dans ce cas-là est de un pour cent par an en moyenne. Il serait intéressant de pousser davantage cette étude afin de déterminer si la divergence se répartit de façon uniforme, ou si elle caractérise une marchandise en particulier.

Quelles qu'en soient les causes, ces comparaisons semblent indiquer que sur un nombre suffisamment grand d'années et pour des parties relativement importantes du total des flux commerciaux, les différences des taux de croissance ne sont pas élevées en termes absolus. Néanmoins, même de petites différences peuvent bouleverser les variations d'une période à l'autre de la balance commerciale globale, en particulier lorsqu'elle est proche de zéro. De plus, lorsque l'on traite avec un partenaire commercial comme le Japon, dont les exportations sont concentrées principalement dans une ou deux catégories à un chiffre de la classification des marchandises, les possibilités de corriger une mauvaise classification systématique sont comparativement peu nombreuses. Il est donc d'autant plus important de comprendre pourquoi le commerce bilatéral, mesuré par les deux dossiers homologues, n'a pas suivi le mouvement.

Un autre type d'analyse est également très révélateur. Tout flux d'importations devrait être égal au flux d'exportations correspondant, plus le coût du fret et de l'assurance, plus un terme qui traduirait l'ensemble des différences de concepts, de datation et des erreurs. Alors que la datation et les erreurs se font surtout sentir à court terme, les différences de concepts devraient se révéler être la cause principale à plus long terme. Pour cette raison, si le rapport des exportations annuelles aux importations annuelles varie dans le temps, cela peut s'expliquer par une combinaison des facteurs suivants: une variation des parts des composants avec un c.f. relativement élevé par rapport à celles avec un c.f. bas, un changement dans la composition des marchandises avec de petites différences quant aux marchandises et d'importantes différences de datation, un changement dans la proportion du c.a.f. à la valeur totale et d'autres facteurs. Le tableau 3 présente quelques résultats agrégés de cette analyse. En regard de chacun des flux touchant le Japon, la CEE et l'Amérique du Nord, il y a trois chiffres: le rapport simple (à pondération de l'année courante) de l'agrégat des exportations à celui des importations en 1978, le rapport correspondant en 1985 et le rapport normalisé pondéré selon l'année de base.

Tableau 1
Différences des taux de croissance annuels du commerce total correspondant pour le Japon, l'Amérique du Nord et la CEE, 1978-1985

Pays A - Pays B	Différences des taux de croissance pour la période ¹			Différence exprimée en valeur des exportations en 1982/arrondie au cinq millions de dollars le plus proche
	1982/78	1985/82	1985/78	

A.N. - CEE	.6	-.5	-.5	265
A.N. - Japon	-.4	.5	-	-
CEE - A.N.	-.8	-.7	-.8	365
CEE - Japon	1.1	1.9	1.5	90
Japon - A.N.	-.7	-.2	-.5	200
Japon - CEE	-1.2	-.6	-.9	155
Moyenne	.8	.7	.7	

¹ Différence définie comme (taux de croissance en % de A^XB) - (taux de croissance en % de $B^{m}A$).
² Différence entre A^XB et $B^{m}A$ arrondie au cinq millions de dollars le plus proche.
 Le symbole (-) indique une valeur insignifiante.

Tableau 2

Différences des taux de croissance annuels du commerce total correspondant par section CTCTI Japon, en 1978-82 et 1982-85

Japon - Amérique du Nord		Japon - CEE	
Section de la CTCTI	Taux de croissance annuel en % pour la période ¹	Différence de la valeur des exportations en 1982	Différence de croissance annuel en % pour la période
	1982/78 1985/82	en 1982	1982/78 1985/82

5. Produits chimiques	.7	-1.6	15	-1.5	.5
6. Semi-produits	-2.5	.9	60	1.9	-.5
7. Matériel de transport	-1.0	-1.0	275	-2.0	-.7
8. Articles manufacturés	1.4	-.9	35	-.8	.8
					20

¹ Différence définie comme (taux de croissance en % de A^XB) - (taux de croissance en % de $B^{m}A$) (A est le Japon).
² Différence entre A^XB et $B^{m}A$ arrondie au cinq millions de dollars le plus proche.

à publier des données bien au-delà du niveau à trois chiffres de la CTCTI ou son équivalent. Un certain nombre de pays du Tiers monde publient des données ventilées selon dix chiffres (classification internationale adaptée au pays) et selon le pays. L'étude permet cependant de constater que les flux codés à un chiffre, où il y a rarement eu des controverses, se caractérisent par des différences très considérables lorsqu'on les compare à leurs homologues des que leur valeur absolue tombe, par exemple, en-dessous de 50 millions de dollars. Au-delà du premier chiffre de la classification, les différences s'accroissent très rapidement.

Le cas des exportations japonaises vers l'Amérique du Nord et les importations correspondantes figurant aux tableaux 1 et 2 mérite qu'on l'examine de plus près. Au point central (1982), ce commerce était évalué à environ quarante milliards de dollars (E.-U.). Le total des importations a augmenté en moyenne de 0.5% par an de plus que les exportations. C'est un montant d'environ 200 millions de dollars par an, pour l'année centrale. Une étude détaillée permet de conclure qu'une partie appréciable de l'explication est attribuable à la section 7 de la CTCTI,

Si les Pays-Bas ne desservent les autres membres de la CEE que comme un port, la création d'un total pour la CEE suffirait à améliorer les comparaisons. Mais d'autres pays, en particulier la Suisse et l'Autriche, bénéficieraient également des ports et des terminaux à conteneurs hollandais. Ceci complique quelque peu le problème, puisque, dans le cas de la Suisse, l'importateur peut utiliser la règle de l'origine dans le cas des Pays-Bas lorsqu'il y a une consolidation des importations provenant d'un grand nombre d'origines, ou bien une transaction de valeur ajoutée à l'extérieur de la zone douanière de Rotterdam peut ne pas être déclarée comme du commerce extérieur hollandais de marchandises.

Un autre obstacle à l'interprétation est posé par les deux Allemands, puisque l'absence tient de déclarer ses importations au BSNL, tandis que l'autre ne considère pas comme des exportations les transactions qu'elle a avec son homologue oriental. Cela veut dire qu'il y a des exportations supplémentaires de la CEE qui n'ont pas de dossiers d'importation, et plus particulièrement, qu'il y a des transactions commerciales non déclarées entre les deux Allemands. L'ampleur de cette fuite non déclarée fluctue selon la prospérité relative de l'Allemagne de l'Est et ne peut être évaluée que par l'étude d'autres indicateurs. Il y a également des fuites pour le commerce avec le Japon et qui touchent les résultats des comparaisons faisant intervenir le Japon et ses partenaires commerciaux. Elles peuvent s'expliquer par des opérations faisant intervenir des succursales de firmes étrangères situées en Asie du Sud-Est. Cependant, l'effet sur les données globales ne devrait pas être important et ne devrait pas reléguer au deuxième plan la valeur de l'analyse utilisant cette base de données.

1) Comparaison des taux de croissance des statistiques correspondantes

Parmi les diverses analyses effectuées à partir de la mini-base de données, une a porté sur la comparaison des taux de croissance des statistiques correspondantes pour la période 1978-1985. On a supposé qu'au cours de cette période, l'effet des erreurs et des différences de données serait suffisamment réduit ce qui ferait ressortir les effets plus permanents. De plus, par l'étude des taux de croissance, il serait possible d'éviter dans une large mesure l'effet d'évaluations différentes. Il est peu probable que la variation du coût de l'assurance et du fret soit suffisamment différente de la variation des prix moyens des marchandises transportées pour modifier sensiblement les taux de croissance sur une période de trois ou de quatre ans. Au moins dans le cas des produits manufacturés, la proportion du fret et de l'assurance dans le coût total est bien inférieure à 10%, comme le montrent les rapports f.o.b./c.a.f. des Etats-Unis. De plus, les coûts du transport ne se rapporteraient qu'au poids et au volume des marchandises transportées. Les coûts d'assurance, qui se rattachent à la valeur, ne représentent pas une proportion appréciable du total des coûts. La substitution du mode de transport a peu de chances de s'ajouter au coût total, sauf dans des circonstances exceptionnelles. Par conséquent, si la variation du coût correspondant était suffisante pour toucher les taux de croissance des importations par rapport aux taux correspondants des importations, les effets seraient tous à sens unique, et leur importance varierait selon la masse moyenne des marchandises transportées. Ces considérations ne sont qu'en partie confirmées par les faits. Le tableau 1 montre les différences des taux de croissance annuels pour le total du commerce correspondant pour des groupes d'origines et de destinations obtenues à partir du commerce entre les pays de la CEE, l'Amérique du Nord et le Japon. Même si ces différences sont relativement faibles, elles ne semblent décaler aucune tendance, malgré quelques régularités éventuelles sous-jacentes échappant à une inspection superficielle.

Le tableau 2 présente les taux de croissance de certaines sections de la CTCl entre le Japon et ses deux partenaires commerciaux. Pour simplifier le tableau, on a posé comme principe d'ignorer les flux inférieurs à un milliard de dollars en 1982, puisque de tels flux ne semblent pas être suffisamment stables pour justifier une interprétation.

Les participants aux discussions portant sur les classifications de marchandises comparables sur le plan international ont inmanquablement exigé davantage de détails, plutôt que moins. La collecte de statistiques aux fins d'une comparaison internationale a poussé les pays

où $A^m B(k)$ est le flux d'importations de la marchandise k du pays B au pays A , enregistré par le pays A ; $B^x A(k)$ est le flux correspondant déclaré par le pays B ; $A(c.a.f.)$ $B(k)$ est l'estimation des coûts de transport et d'assurance pour ce flux commercial, calculée à partir des dossiers du pays A ; θ est un ajustement de période et e un terme d'erreur qui comprend tous les biais et erreurs aléatoires qui caractérisent la statistique des importations et celle des exportations. On suppose que toutes les autres causes de différence (géographie, inclusions et exclusions, livraisons de faible valeur, etc.) ont été réglées soit par correction, ou, de préférence, par l'exclusion de toutes les transactions qui pourraient être entachées de ces facteurs des fichiers de comparaison. Dans le temps, l'erreur moyenne devrait tendre vers zéro, et par conséquent, plus la période au cours de laquelle la comparaison est faite est longue, et plus proche est le niveau du taux moyen de variation des chiffres comparés. Si une telle comparaison devait soudainement donner des résultats pervers, on aurait alors la preuve évidente d'une détérioration de la qualité d'au moins un des deux termes de la comparaison.

3.1 Analyse utilisant la mini-base de données du commerce mondial

À des fins d'analyse, on a créé une mini-base de données à partir des données du commerce mondial afin d'entreprendre l'étude de quelques-uns de ces effets. Elle recouvre les trois principaux blocs commerciaux du monde occidental: la CEE, définie dans ce cas par l'exclusion du Portugal et de l'Espagne, l'Amérique du Nord (Canada et États-Unis) et le Japon. En plus d'être plus simple d'utilisation en raison du nombre réduit d'enregistrements, elle évite le problème des déclarations en retard (principalement les pays du Tiers monde) et de la non-déclaration (principalement les pays à économie centralisée). La mini-base de données comprend les données sur les exportations et les importations de chaque pays, ventilées par la CTCT (jusqu'au niveau de détail à quatre chiffres) et par pays partenaire de 1978 à 1985. En plus des pays participants, elle comprend deux agrégats, la CEE et l'Amérique du Nord. Alors que la base de données du commerce mondial utilise un certain nombre d'imputations pour simplifier l'analyse, la mini-base ne comprend que les données que les pays membres ont déclarées au BSNL, après que ce dernier eût regroupé des catégories commerciales considérées secrètes par le pays déclarant et converti les codes non normalisés déclarés par les pays en codes normalisés CTCT. Aucune de ces transformations ne devrait affecter sensiblement les conclusions résultant de la base de données.

Le groupement des pays dans la mini-base de données pose quelques problèmes statistiques. Les États-Unis déclarent leurs importations au BSNL sur la base c.a.f., mais le Canada déclare ses importations f.o.b. Alors que les États-Unis créditent leurs pays partenaires en fonction de l'origine des biens importés, le Canada les déclare selon le pays de consignation, à l'exception des importations provenant des pays d'Amérique latine. Ceci ne serait pas un problème trop grave s'il n'y avait le fait que les États-Unis sont parfois crédités pour des exportations se dirigeant vers le Canada. Par conséquent, bien que l'addition des deux pays devrait améliorer l'appariement des flux correspondants, les différents systèmes d'enregistrement le rendent beaucoup plus difficile. On espère que cet inconvénient disparaîtra lorsque les importations f.o.b. des États-Unis seront versées dans la base et lorsque les importations canadiennes selon l'origine remplaceront les importations selon la consignation pour le plus grand nombre d'années antérieures possibles. Dans le cas des pays de la CEE, le rôle essentiel des Pays-Bas comme point d'entrée pour le continent européen rend les comparaisons difficiles. La zone douanière du port de Rotterdam agit non seulement comme un centre de distribution géant, mais également comme un entrepôt pour les pays desservis. L'exportateur à l'extérieur de la CEE peut donc ne pas donc connaître à quel pays en particulier la vente est faite, et tout ce qu'il saura c'est que l'entrepôt aura lieu à Rotterdam; pour cette raison, il créditera les ventes aux Pays-Bas. Mais l'importateur ultime est lié par la règle de l'origine pour attribuer l'achat au bon pays. Dans le cas des Pays-Bas, selon leurs dossiers, aucune transaction portant sur des biens avec franchissement de leurs frontières douanières n'a eu lieu. Le pays a simplement vendu des services portuaires et d'entreposage à l'une des parties.

3. UN PROGRAMME DE MESURE DES ERREURS

Les causes des erreurs sont connues depuis longtemps (Coats 1926). Au début des années 70, une tentative valable de quantification a été faite lors du premier projet de rapprochement entre les Etats-Unis et le Canada. Mais jusqu'à présent, c'est une très importante proportion de ce que l'on connaît des erreurs dans les statistiques du commerce extérieur, et il est évident qu'elle se trouve limitée par le fait qu'il s'agit du commerce entre deux pays limitrophes, et entre ces deux pays seulement. Etant donné que les bases de données internationales comme celle du Canada vont probablement devenir plus répandues et qu'elles utiliseront divers logiciels analytiques, le moment est venu de s'interroger sur ce que l'on pourrait faire pour améliorer la statistique du commerce, ou en l'absence de toute amélioration, au moins pour informer les utilisateurs des limites des données du commerce extérieur. Il est actuellement peu probable que l'on tienne compte des renseignements descriptifs disponibles, que les utilisateurs d'un pays vont se rendre compte du degré d'erreur des tendances à long terme de la statistique du commerce, ou de la mesure dans laquelle les mouvements mensuels de leurs balances commerciales nationales respectives se trouvent affectés, et, ce qui est plus important encore, dans quelle mesure les renseignements au niveau détaillé de marchandises sont susceptibles de contenir des erreurs. Il est évident que le flux A_{XB} pourrait être le même que B_{mA} , tant que toutes les livraisons et leur enregistrement sont instantanés, que la base d'évaluation est la même pour les deux par-tenaires pour la même transaction, que les règles d'inclusion et d'exclusion sont les mêmes, qu'il n'y a pas de différences conceptuelles (géographiques, comptables ou résultant des dis-positions douanières) et qu'il n'y a pas d'erreurs (de codage ou de champ). On inclut dans les «erreurs» les interprétations cohérentes des protocoles de classification d'un pays qui pour-raient être contestés par d'autres pays ou par le Conseil de coopération douanière.

En principe, toutes les causes de différences autres que des erreurs devraient être résolubles, bien que la mesure de l'importance relative des différentes sources puisse être difficile dans la pratique. Une étude des différentes causes ou des facteurs est utile si l'on veut étudier la façon dont leur effet peut être pris en compte dans toute comparaison. De ces facteurs, le transport est probablement le moins difficile à traiter et presque certainement, le moins difficile à aborder. Il y a un certain nombre de pays comme les Etats-Unis, où les importations sont mesurées de deux façons, avec et sans le transport. En principe, les renseignements servant à estimer le coût de l'assurance et du fret existent habituellement. Les importateurs doivent, en vertu de la loi, informer leurs autorités douanières de toutes leurs dépenses au titre d'un achat à l'étranger, et les deux catégories générales de dépenses sont celles qui sont impossibles (habituellement, les dépenses reliées au produit proprement dit, y compris l'emballage, ou les frais ou la monture) et toutes les autres (y compris celles reliées au transport, à l'assurance et au financement de l'importation). Par conséquent, s'il fallait exécuter une étude des coûts de transport, il y aurait des dossiers administratifs qui pourraient être reliés aux dossiers correspondants du commerce. Il y a un grand nombre de problèmes techniques reliés à la façon dont les renseignements sur le fret et l'assurance pourraient être attribués à chaque produit dans le cas des livraisons com-plexes, mais il y a des suggestions pour traiter ces problèmes (Ryten 1983).

On pourrait également en principe entreprendre une étude des différences de datation dans le contexte d'un flux commercial particulier entre deux pays. Dans le cas du rapprochement de la statistique du commerce entre les Etats-Unis et le Canada, les estimations ont utilisé des rapprochements réels de documents, ce qui a permis de comparer les dates et les délais moyens entre les exportations et les importations correspondantes. Mais il y a des méthodes moins oné-reuses pour obtenir des estimations approximatives qui sont également moins limitées sur le plan de l'accès à des dossiers confidentiels et qui sont suffisamment efficaces pour permettre le calcul d'intervalle généraux de différences de datation par point de sortie et d'entrée, par mode de transport et par produit.

Ensemble, les estimations des différences de datation et la différence entre le coût de l'assu-rance et du fret et les évaluations franco à bord peuvent être exprimées sous la forme suivante:

$$A_m B(k) = B_{XA}(k) + A(c.a.f.) B(k) + \theta + e$$

de services élevée tels que les bandes audio et vidéo enregistrées, les plans des architectes, les logiciels enregistrés sur bandes magnétiques, les réparations et l'entretien, etc.

v) *Différences entre les dossiers des exportations et des importations: le codage et le traitement des données*

Pratiquement toutes les catégories d'information qui figurent sur les dossiers principaux tenus par les Douanes traduisent l'application d'une classification ou d'un code à une situation réelle. La cohérence du codage peut être garantie par des décisions sur les cas limites et en veillant à ce que ces décisions prises ensemble constituent en quelque sorte une jurisprudence, c'est-à-dire un ensemble de décisions à prendre, accessibles aux coders et qui régiraient ces derniers. Mais le seul organisme central réglementaire est le Secrétaire du Conseil de coopération douanière, qui se trouve à Bruxelles. Les pays membres ne peuvent s'y adresser sur une base quotidienne, et ses décisions ne peuvent dépasser un certain niveau de généralité. Pour cette raison, il y a des différences systématiques dans l'interprétation et l'application de codes normalisés parfois à l'intérieur d'un même pays, et à plus forte raison entre pays.

Il y a par ailleurs des incohérences imputables aux erreurs au moment du traitement des données et aux systèmes mis en place pour réduire leur impact. Ainsi, il y a des erreurs d'interprétation de la législation douanière et lors du codage des sources de renseignements qui se glissent au moment où les importateurs ou les exportateurs informent leurs autorités respectives d'une livraison imminente, des erreurs au moment de la saisie des données et des erreurs de codage au Bureau de statistique. La protection habituelle contre ces erreurs est la mise en place de systèmes de revue et de contrôle qui utilisent à des degrés différents une inspection et une revue manuelles et la détection et l'imputation informatiques. Même s'il est très probable qu'il y ait d'autres causes d'incohérences, celles mentionnées ci-dessus sont les plus fréquemment citées, depuis qu'elles ont été décrites pour la première fois (Coats 1926), et elles sont probablement les plus importantes dans le cas des différences entre les chiffres correspondants.

vi) *Différences entre les dossiers des exportations et des importations: les quantités, variable spéciale*

Contrairement aux valeurs, les quantités déclarées ne sont pas touchées par l'inclusion des coûts de transport et elles ne sont pas biaisées afin de minimiser les créances fiscales, cependant, si des valeurs sont mal codées dans les catégories à droits plus faibles, les quantités correspondantes vont suivre. Malheureusement, il y a d'autres problèmes associés à l'enregistrement et à l'utilisation des quantités qui réduisent considérablement la valeur de ces statistiques lors de la détection des erreurs. Ainsi, les quantités peuvent s'appliquer soit à une livraison complète, auquel cas elles sont habituellement exprimées comme un poids brut, soit à une marchandise en particulier, auquel cas elles sont exprimées soit en poids nets, ou en toute autre unité appropriée (longueur, surface, volume) y compris, dans le cas des marchandises complexes, le nombre d'unités.

Alors que les mesures de la quantité, en poids bruts ou nets sont comparables entre pays, leur utilisation se trouve limitée par l'hétérogénéité des livraisons en cause. Les quantités exprimées en d'autres unités se trouvent limitées par la diversité des unités utilisées, et ce qui est plus important encore, par le fait qu'elles ne peuvent être agrégées dans la classification par marchandises, et les niveaux auxquels elles s'appliquent sont beaucoup trop détaillés pour une comparaison entre pays compte tenu de nos connaissances actuelles. Ces unités trouvent également une utilisation dans la comparaison du commerce des matières brutes, en particulier si, avec les valeurs, on les utilise pour suivre les fluctuations des valeurs unitaires. En fait, a proposé à la dix-huitième session de la Commission statistique des Nations Unies (Rapport du Secrétaire Général 1974) une étude internationale des erreurs de commerce utilisant principalement la comparaison des valeurs unitaires. Cependant, les pays membres n'ont pas estimé que les avantages éventuels justifiaient le coût prévu. À l'heure actuelle, la base de données du commerce mondial de Statistique Canada n'inclut pas des données sur les quantités, de sorte que les applications de la statistique des quantités n'ont pas encore été étudiées.

Les différentes lettres signifient que certains pays déclarants imputent leurs exportations au premier pays de destination connue, et d'autres, au dernier pays de destination connue, que certains pays importateurs imputent les importations au pays d'origine et d'autres au pays de consignation, et que certains pays exportateurs comptent comme exportations tout ce qui sort de leur territoire national, quel que soit le degré de transformation des biens en cause. Les différencés de ces approches ne sont pas négligeables à une époque où l'on parle d'accords de libre-échange, d'union douanière, de zones de libre-échange et d'autres accords pour stimuler le commerce transfrontalier. Dans chaque cas, une convention statistique distincte est nécessaire pour tenir compte de l'effet de l'accord sur la comptabilisation douanière. Créditer les pays partenaires de façon incohérente n'est qu'une cause de divergence dans les comparaisons bilatérales ou multilatérales. L'autre est due au manque d'uniformité de la classification géographique. En fait, un grand nombre de pays incorporent leur position en matière de politique extérieure dans leurs classifications géographiques types. Par conséquent, des définitions géographiques non cohérentes des pays partenaires entraînent des différences. La plupart des pays d'Amérique latine traitent Porto Rico comme une origine ou une destination distincte des États-Unis. Pratiquement chaque pays membre de l'OCCDE traite différemment les pays partenaires d'Afrique. Certains les regroupent selon leurs origines coloniales, tandis que d'autres le font selon le voisinage géographique. Des incohérences semblables se produisent dans le cas des îles des Caraïbes et du Pacifique sud. L'union économique d'Afrique du Sud est traitée dans les statistiques d'une façon qui traduit souvent l'opinion du pays déclarant d'un embargo sur les relations commerciales avec l'Afrique du Sud proprement dite. De plus, tous les pays n'engrèstent pas les changements de statut politique de leurs partenaires commerciaux avec le même zèle, et pour cette raison tous les pays ne suivent pas les nouveaux pays indépendants aussi rapidement que cela serait souhaitable pour effectuer des comparaisons statistiques.

(iv) *Différences entre les dossiers des exportations et des importations: l'administration douanière*

Une autre différence importante surgit du fait que les administrations douanières accordent une attention moindre aux exportations que leur mandat ne l'exige pour les importations. La déclaration des livraisons d'exportations individuelles peut être consolidée pour des considérations de réduction de la paperasserie et alignée sur les bordereaux d'expédition ou les autres documents de transport des transporteurs. Dans les cas des importations, le but visé est d'obtenir une déclaration suffisamment détaillée pour permettre aux Douanes de prélever les droits et autres taxes corrects. Une conséquence est que, dans le cas des exportations, les articles de faible valeur d'une livraison complexe auront probablement plus de chance d'être classés sous la même rubrique que les principales composantes, alors que dans le cas des importations, ils risquent d'être classés de façon indépendante.

Cette différence d'intérêt que l'on peut attribuer au mandat de l'administration douanière a d'autres conséquences importantes sur la qualité des documents d'exportation et d'importation. D'une part, il est évident que l'ampleur de la sous-déclaration des exportations qui caractérise les exportations terrestres des États-Unis vers le Canada ne se limite pas à l'Amérique du Nord. Il y a presque vingt ans, le Royaume-Uni a entrepris un vaste programme qui consistait à rapprocher les bordereaux d'expédition aux documents d'exportation à cause d'un taux de sous-déclaration soupçonné d'environ un à deux pour cent du total. Par contre, on estime que la description des produits exportés est non-biaisée, à moins qu'il ne s'agisse de livraisons illégales, tandis que la description des biens importés pourrait être faussée dans le but de minimiser les taux des droits impossibles.

En plus de ces raisons, qui s'expliquent par les différentes transformations juridiques et administratives que subit le dossier original, il y en a d'autres qui sont plus variables et plus sélectives quant aux dossiers visés. On peut citer comme exemples le traitement des livraisons de faible valeur, qui sont définies comme étant inférieures à différents seuils exclus, inclus ou échantillonnées à des taux différents, et le traitement des produits qui contiennent une proportion

!!) Différences entre les dossiers des exportations et des importations: les valeurs

Les différences de valeurs ont pendant longtemps empêché des comparaisons systématiques, et il est bon de les revoir et d'évaluer leur importance relative. L'évaluation de la transaction, c'est-à-dire le prix auquel elle est enregistrée aux fins douanières, est essentielle. Un grand nombre de pays, sinon la plupart, enregistrent la valeur d'une importation en y incluant le coût du transport international et de l'assurance rattaché à la livraison. La plupart des pays enregistrent la valeur des exportations correspondantes sans ces éléments. Il y a d'autres écarts: certains pays incluent des éléments des coûts du transport intérieur et de l'assurance, tandis que d'autres excluent les coûts portuaires du transport international. Mais ces différences ne représentent qu'une augmentation minimale de la difficulté de comparer les dossiers correspondants. Les transactions faisant intervenir des partenaires commerciaux reliés, comme dans le cas des entreprises multinationales, posent un problème d'évaluation qui est résolu différemment selon les pays. Il est possible que cette cause de différence finisse par supplanter toutes les autres au cours des années à venir.

!!!!) Différences entre les dossiers des exportations et des importations: le pays

La question du pays créditeur peut introduire quelques-unes des plus bizarres différences dans tout programme systématique de comparaison. En tant qu'exportateur, un pays peut considérer comme exportation toute vente de produit qui doit franchir ses frontières douanières pour atteindre sa destination, indépendamment du fait que ce produit a été sensiblement transformé ou vendu exactement en l'état qu'il a été acheté auprès d'un autre pays. Cependant, comme importateur, un pays peut décider d'imputer un achat au pays où la dernière transformation appréciable (habituellement "appréciable" à une définition précise selon la loi) a eu lieu. Par conséquent, si l'on prend le cas de trois pays hypothétiques A, B et C, où A a exporté des biens à B et B a exporté les mêmes biens (peut-être après transformation) à C, les statistiques peuvent être enregistrées de façons très diverses avec des conséquences différentes, comme le montre le tableau ci-dessous. Les symboles x et m dénotent respectivement la valeur des exportations vers le pays partenaire et des importations de celui-ci (deuxième lettre majuscule) (première lettre minuscule).

Nous avons donc:

$A^x B$ = Valeur des exportations de A à B enregistrée par A
 $A^m B$ = Valeur des importations de B à A enregistrée par A

	Enregistrées comme exportations	Enregistrées comme importations	Conséquence
i)	$A^x B + B^x C$	$B^m A + C^m B$	Cohérence et exhaustivité
ii)	$A^x B + B^x C$	$B^m A + C^m A$	Sur-imputation de A par les importateurs
iii)	$A^x C + B^x C$	$B^m A + C^m B$	Sur-imputation de C par les exportateurs
iv)	$A^x C$	$C^m A$	Cohérence mais non-exhaustivité
v)	$A^x C$	$C^m B$	Pas d'imputation à A par les importateurs
vi)	$A^x B$	$C^m A$	Pas d'imputation à C par les exportateurs.

la transaction, qui comprennent une description du, ou des, produits vendus, la valeur et la quantité qui y correspondent, les conditions de la vente, une identification de l'acheteur et de la résidence de celui-ci, ainsi qu'une date à laquelle la transaction a eu lieu ou aura lieu. Ce dossier produit un certain nombre de dossiers connexes, les uns obtenus par la transformation des renseignements de base sous une certaine forme prescrite, les autres, par le raccordement de ces dossiers à des dossiers connexes. On peut citer comme exemples de ce dernier cas une description de la façon dont les produits échangés passent du lieu de vente au lieu d'achat et le coût de cette opération, le coût de l'assurance de la livraison, les montants à facturer aux deux parties à la transaction à cause des droits, les taxes de vente, les frais consulaires, etc., et naturellement, la forme et la date de règlement de l'achat.

Les transformations de ces renseignements de base se rattachent aux conventions touchant la façon dont ces renseignements sont enregistrés et la documentation des différentes étapes de la transaction dans le temps. Ces transformations ne sont pas uniformes d'un pays à l'autre. Les conventions qui les régissent sont soit exposées dans la législation douanière, soit dans les règlements administratifs qui régissent la tenue des dossiers douaniers. Ils donnent lieu aux documents qui constituent la base de la statistique du commerce extérieur. Le pays de vente conserve un ensemble de documents, le pays d'achat, l'autre. Dans la pratique, ces documents vont différer en dépit du fait qu'ils se rattachent en principe et dans les faits à la même transaction commerciale. D'abord, ces documents diffèrent dans le temps. Même dans le cas de pays limitrophes, ou lorsque le transport aérien intervient, les différences temporelles ne sont pas négligeables. Elles surrissent parce que l'ensemble des liens qui constituent la transaction est long, puisqu'il s'agit de transporter la marchandise au point de départ du transporteur international, de l'entreposer avant le départ du transporteur international, de l'amener au point de destination, de l'entreposer en attendant de remplir les formalités douanières, et pendant toutes ces diverses étapes, de remplir les documents à différentes étapes et les enregistrer en fonction des différentes conventions en vigueur.

De plus, dans un pays la date de la transaction peut être inscrite comme étant celle au moment où la facture est reçue dans le pays importateur, et dans un autre, au moment où les autorités douanières perçoivent les droits. Dans un pays, l'enregistrement de la valeur de l'achat peut comprendre tous les coûts du transport international et de l'assurance, alors que dans un autre, ces coûts peuvent être distincts. Ensuite, les unités de déclaration des quantités peuvent entraîner des incohérences. Par ailleurs, dans un pays la transaction peut être imputée non pas au pays où la facture a été émise, mais plutôt à celui où le produit a été cultivé, extrait ou fabriqué, tandis que dans un autre, ce sera le lieu de résidence du vendeur qui décidera de l'identification du pays. Les situations politiques peuvent également toucher la façon dont un pays est identifié dans les dossiers. Les règlements douaniers peuvent introduire un biais dans la façon dont les importations ou les exportations sont enregistrées. Enfin, il y a des erreurs de codage et de traitement des données. Dans les sections qui suivent, on trouvera d'autres renseignements sur ces divers facteurs.

i) Différences entre les dossiers des exportations et des importations: la datation

Les autorités douanières vont habituellement classer les dossiers de diverses façons: selon le pays d'origine, selon l'identification de l'entreprise importatrice ou de son agent et selon la date de réception. Mais toute opération d'importation comporte au moins quatre éléments fondamentaux, qui peuvent tous être enregistrés, mais dont un seul sera retenu comme date pour l'extraction et les statistiques. Le choix de la date ne fait pas l'objet d'une normalisation statistique, mais dépend plutôt de la façon dont les Douanes considèrent leur fonction principale et de la capacité technique de stocker d'autres données. Il est évident que si un pays considère comme date d'enregistrement des exportations le moment auquel l'organisme expéditeur prépare un document d'exportation, et si le pays d'importation choisit comme date d'importation celle à laquelle tous les droits et autres frais doivent être réglés, l'écart possible entre l'enregistrement des exportations et celui des importations correspondantes atteint un maximum.

2. LES DOSSIERS DES OPÉRATIONS COMMERCIALES: ERREURS ET DIFFÉRENCES DANS LES DOSSIERS CORRESPONDANTS

des très rares où il est possible de procéder à une comparaison de deux mesures de la même transaction calculée pratiquement au même niveau de détail avec les mêmes méthodes par deux enquêteurs indépendants. Les différences qui résultent lors de l'établissement de ces comparaisons ont été mentionnées dans la littérature consacrée à ce sujet, qui remonte au moins jusqu'à la première guerre mondiale (Coats 1926). Mais ces études n'ont pas donné lieu à des projets d'incorporation des résultats de ces comparaisons dans tout document sur la qualité des statistiques. Un des obstacles à des comparaisons systématiques de ce genre pourrait être la quantité de calculs en cause et les coûts. Un autre pourrait être le degré de connaissance nécessaire des systèmes statistiques correspondants qui, en plus d'être présentes dans certains cas dans une langue étrangère, comprennent habituellement des dispositions administratives et légales très précises qui ne sont pas comparables d'un pays à l'autre.

Les obstacles à des comparaisons systématiques ont été surmontés dans une certaine mesure à Statistique Canada, où l'on a établi une base de données du commerce mondial. Cette banque contient des statistiques détaillées du commerce des pays qui déclarent des données sous une forme lisible par une machine au Bureau de statistique des Nations Unies (BSNU). Les pays membres des Nations Unies conviennent, en vertu des conditions d'appartenance à l'Organisation, de déclarer un certain nombre de statistiques-clés au BSNU de la manière que le Secrétaire général des Nations Unies le précise. Ces statistiques comprennent celles du commerce extérieur, ventilées par pays et par produit, et dans ce dernier cas, de façon complète grâce à la Classification type pour le commerce international (CTCI) ou son équivalent, la Nomenclature du Conseil de coopération douanière (NCCD). Les rapports annuels sous une forme lisible par une machine remontent jusqu'au début des années 60.

La base de données du commerce mondial a été créée afin d'aider les négociateurs canadiens qui participent à la série actuelle des réductions multilatérales des tarifs, et aussi pour aider les exportateurs et les importateurs canadiens à mieux comprendre les marchés et les fournisseurs avec lesquels ils traitent. L'inconvénient de la base de données est qu'elle n'est pas complète. Les économies à planification centralisée ne déclarent aucune donnée ou ne fournissent que des données très agrégées. Un grand nombre des pays du Tiers monde accusent de graves retards pour le traitement de leurs dossiers douaniers, ce qui signifie que beaucoup de données manquent pour les années récentes. Tous les pays n'utilisent pas la même édition de la CTCI, et il existe encore pas mal de différences dans les concepts et les définitions utilisés par les différents pays.

Mais ces inconvénients sont neutralisés par le fait que les calculs en cause pour la comparaison de la statistique du commerce sont maintenant plus pratiques, qu'une très importante proportion du commerce mondial ne touche que les pays occidentaux et qu'elle fait l'objet d'une déclaration courante, laquelle utilise progressivement des concepts plus comparables. Si l'on prend en compte tous ces éléments, une base de données du commerce mondial pourrait servir à présenter les résultats d'une comparaison de la statistique commerciale correspondante, ce qui, à son tour, pourrait aider les organismes statistiques à s'intéresser davantage aux forces et aux faiblesses de leurs données sur les importations et les exportations de marchandises. C'est une condition nécessaire pour améliorer la fiabilité de la statistique du commerce. Compte tenu de l'attention qui est actuellement accordée à ces données, les organismes statistiques du monde entier feraient bien d'apporter les améliorations nécessaires que révèlent les comparaisons bilatérales des données correspondantes, même si cela n'est possible que de façon progressive. Dans les sections suivantes, on examine les principales causes de divergence de la statistique correspondante et les mesures qui pourraient être prises pour estimer leur importance relative dans des cas particuliers.

À la base de deux dossiers commerciaux correspondants, on retrouve dans la plupart des cas une seule transaction documentée. Un exportateur a conclu une vente et facturé l'acheteur en conséquence. Cette facture va probablement contenir des renseignements essentiels sur

Les erreurs dans les statistiques du commerce extérieur

JACOB RYTEN¹

RÉSUMÉ

Malgré la facilité relative de l'étude des erreurs dans les statistiques du commerce extérieur, il y a eu peu de tentatives pour quantifier leur taille, leur origine, leur distribution et leur évolution dans le temps. Les décideurs et les négociateurs commerciaux n'ont que des notions très limitées de l'imprécision de ces statistiques, en dépit de leur niveau de détail poussé. L'auteur s'est servi de la banque de données du commerce mondial mise au point à Statistique Canada pour étudier et quantifier les divergences qui existent dans les statistiques du commerce extérieur.

MOTS CLÉS: Commerce extérieur; balances du commerce bilatéral; erreurs.

1. INTRODUCTION

On examine ici quelques-unes des causes d'erreurs dans la statistique du commerce extérieur, les difficultés que pose leur détection, les façons de réduire l'imprécision des chiffres détaillés et un projet d'amélioration de la qualité des données.

Depuis l'ouvrage d'Allen et Ely (1953), publié il y a trente-cinq ans, la question des erreurs dans la statistique du commerce extérieur n'a guère soulevé l'intérêt des auteurs. On s'est intéressé quelque peu aux questions comparables, c'est-à-dire les inclusions et les exclusions, la définition des limites, l'évaluation, etc. (Nations Unies 1982), mais principalement, à la classification. En fait, l'un des plus importants changements apportés à la classification du commerce vient tout juste d'être mis en pratique (Nations Unies 1986) afin de rendre les données du commerce extérieur plus comparables entre pays. Mais, peut-être parce que ces statistiques utilisent une comptabilité complète de toutes les transactions sur les marchandises qui franchissent les frontières au cours d'une période donnée et parce que cette comptabilité est appliquée par un organisme réglementaire, à savoir l'administration douanière, on a tendance à penser en général qu'il ne reste plus guère d'erreurs mesurables. L'absence d'une analyse des erreurs de ces statistiques vient confirmer cette idée.

Régulièrement, on constate, en particulier dans le cas des bureaux de statistique des organismes internationaux, qu'il y a une erreur grave dans la comptabilisation du commerce entre deux pays. À sa dix-huitième session, la Commission statistique des Nations Unies (1974) a été officiellement notifiée du rapprochement de la statistique du commerce entre les États-Unis et le Canada. Ce rapprochement était motivé par la détection de quelques différences embarrassantes de la balance du commerce bilatéral entre les deux pays. Ensuite, et à diverses occasions, on a examiné des problèmes mettant en cause Singapour et la Malaisie, Singapour et l'Indonésie et divers pays hors CEE et les Pays-Bas les organismes internationaux qui s'intéressent plus particulièrement aux questions du commerce. De plus, les pays qui estimaient qu'ils perdaient le contrôle de la qualité de leur statistique du commerce extérieur, essentiellement les pays du Tiers monde, ont tenté de reconstituer leurs propres chiffres en se référant à ceux de leurs principaux partenaires commerciaux. Mais rien ne prouve qu'aucune de ces manifestations d'inquiétude se soit traduite par un programme systématique de détection, de mesure et de réduction des erreurs dans les statistiques.

On ne relève que peu d'explications possibles évidentes pour ce manque d'initiative, à part celle de la croyance qu'il n'y avait aucune erreur. La statistique du commerce extérieur est l'une

¹ Jacob Rytén, Statisticien en chef adjoint, Statistique Canada, 13-B8 Immeuble Jean Talon, Tunney's Pasture, Ottawa, Ontario K1A 0T6.

L'auteur explique les causes de divergence dans les statistiques sur le commerce correspondantes et il analyse l'importance relative de chaque cause. D'après les résultats d'une étude sur les données relatives aux importations et aux exportations tirées de la base de données sur le commerce international de Statistique Canada, l'auteur met sérieusement en doute la comparabilité des données sur le commerce correspondantes aux niveaux détaillés de la classification des marchandises. Il propose un programme qui améliorerait la qualité des statistiques sur le commerce extérieur et explique pourquoi il faudrait donner aux utilisateurs des renseignements concrets sur la qualité des données.

En pratique, il arrive souvent que les données tirées de la plupart des enquêtes servent à plusieurs usages. Cependant, les rapports de recherche et les ouvrages statistiques évitent habituellement le sujet des plans de sondage à usages multiples. Kish se penche sur cette question importante dans son article. Il présente pour commencer une hiérarchie de ces usages, puis il parle des exigences contradictoires auxquelles doivent satisfaire les plans de sondage des enquêtes à usages multiples. Il traite de dix sources d'incompatibilité, dont le choix de la taille de l'échantillon et la répartition de celui-ci en domaines et en strates, le rapport entre les biais et les erreurs d'échantillonnage, le choix des variables de stratification et la continuité des données. Pour chacun de ces problèmes, l'auteur propose des solutions et recommande fortement l'utilisation de plans de sondage qui constituent des compromis plutôt que de plans optimaux pour un seul usage. Certaines propositions sont moins rigoureuses que les autres et sont avancées pour stimuler la recherche sur ce sujet.

Dans leur article "Sur la stratification de populations asymétriques", Lavallée et Hidiroglou proposent un algorithme itératif pour la stratification de populations asymétriques dans les cas où l'échantillon est proportionnel à la puissance p (c.-à-d. une répartition de l'échantillon proportionnelle au nombre total d'éléments dans chaque strate élevée à une petite puissance). Ils font une étude empirique dans laquelle la méthode de répartition suggérée est comparée à d'autres méthodes de répartition et où ils utilisent des données tirées des enquêtes annuelles sur le commerce de détail et le commerce de gros menées par Statistique Canada.

La SIPP (Survey of Income and Program Participation/enquête sur le revenu et la participation aux programmes) est une enquête permanente menée par le U.S. Bureau of the Census. Dans son article "Domaines de recherche relatifs à la SIPP", Kasprzyk passe en revue les questions d'ordre méthodologique et statistique que soulève la SIPP. Il se penche sur quatre questions en particulier qui concernent toutes les enquêtes par panel menées auprès de familles et de particuliers. Il s'agit de la conception du questionnaire, de la collecte des données, de l'erreur de réponse et enfin du plan de sondage et de l'estimation pour des variables longitudinales. L'article contient une description des problèmes importants, renvoie à des études où ces problèmes ont été examinés et résume les principaux résultats de ces études.

Dans l'article "Logiciel d'ordinateur personnel pour l'estimation de la variance dans des enquêtes complexes", Schnell, Kennedy, Sullivan, Park et Fuller décrivent un programme appelé PC CARP qui a été conçu pour l'analyse des données tirées d'enquêtes complexes. On a trouvé des applications à ce programme, en particulier dans plusieurs pays en voie de développement. Les caractéristiques et les fonctions du logiciel sont décrites brièvement.

Dans ce numéro

Quatre des neuf articles de ce numéro traitent de l'erreur de couverture dans le recensement. Ces articles, de même que ceux qui paraîtront dans le numéro de décembre 1988 de la revue, viennent enrichir la documentation croissante sur le sujet. L'initiative dont a fait preuve Kirk Wolter a permis de réaliser ces sections spéciales.

Nous savons que les chiffres du recensement sont inexacts à cause de l'erreur de couverture et ces problèmes ont suscité beaucoup d'intérêt ces derniers temps, aussi bien auprès des statisticiens (théoriciens et praticiens) que des décideurs. En conséquence, les méthodes permettant d'évaluer la qualité de ces chiffres de même que les limites de ces méthodes, les techniques de redressement (basées aussi bien sur le plan de sondage que sur des modèles) visant à améliorer la qualité des chiffres de population, l'incidence du sous-dénombrement sur les programmes gouvernementaux et d'autres études du même ordre ont pris une importance grandissante. Dans de nombreux pays, des études d'évaluation sont faites pour estimer l'erreur de couverture soit pendant le recensement, soit après. Ainsi, au Canada, la contre-vérification des dossiers est la plus importante étude réalisée pour mesurer le sous-dénombrement. Aux États-Unis, c'est l'enquête postcensitaire (EP) qui est, depuis 1950, un des principaux moyens utilisés pour mesurer le taux de couverture du recensement.

En 1986, le U.S. Bureau of the Census a mené, à Los Angeles, une étude intitulée "test des opérations de redressement" (TOR) qui mettrait à l'essai un nouveau plan de sondage pour l'EP. Trois articles de la section spéciale, soit ceux de Diffendal, Schenker et Hogan et Wolter, fournissent une évaluation détaillée des méthodes et procédures utilisées dans cette nouvelle EP et analysent à fond les résultats de la recherche ainsi que les difficultés rencontrées pendant le TOR et les résultats obtenus au moyen de ce test. Diffendal donne un aperçu général de celui-ci, décrivant les méthodes utilisées et les opérations exécutées. Il présente également un bref historique des études réalisées aux États-Unis pour estimer le taux de couverture du recensement et décrit les événements récents qui ont amené le U.S. Bureau of the Census à effectuer des études complexes.

Schenker traite de trois méthodes utilisées pour régler le problème des données manquantes: l'imputation par hot deck, l'utilisation des modèles de régression logistique et le redressement par la pondération. Le choix d'une méthode dépend du type de données manquantes. On se sert, par exemple, de la régression logistique pour imputer les valeurs des caractéristiques binaires. Se servant des données du TOR, l'auteur compare les estimations de l'erreur de couverture obtenues au moyen de diverses méthodes d'imputation.

Hogan et Wolter examinent en détail les sources possibles d'erreur dans les nouvelles estimations produites à partir des résultats de l'EP et évaluent l'incidence de chaque type d'erreur ainsi que de l'ensemble des erreurs sur les données du TOR. D'après leurs observations, les auteurs concluent que, en pratique, "il pourrait alors y avoir des secteurs pour lesquels les estimations de l'EP seraient plus justes que les estimations du recensement et d'autres (la plupart de ceux qui restent) pour lesquels elles seraient aussi justes sinon presque aussi justes que les estimations du recensement".

Le quatrième article de la section spéciale, "Modélisation de l'erreur d'appariement et son effet sur les estimations de l'erreur d'observation du recensement", de Biemer, traite du problème d'appariement des données de l'EP et de celles du recensement. L'auteur prend trois modèles de plus en plus complexe et examine l'incidence de l'appariement sur les estimations de l'EP. Il parle ensuite des conséquences de ces résultats pour le recensement de 1990.

Les cinq autres articles dans ce numéro portent sur les erreurs dont sont entachées les statistiques sur le commerce extérieur, les problèmes que soulèvent les plans de sondage des enquêtes à usages multiples, la stratification des populations dissymétriques, l'enquête sur le revenu et la participation aux programmes menée par le U.S. Bureau of the Census et un logiciel d'ordinateur personnel pour l'estimation de la variance.

Dans son article "Les erreurs dans les statistiques du commerce extérieur", Rytien traite des sources d'erreur dans ces statistiques et des méthodes permettant de réduire ces erreurs, et il

TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada
Volume 14, numéro 1, juin 1988

TABLE DES MATIÈRES

Dans ce numéro	1
J. RYTEN Les erreurs dans les statistiques du commerce extérieur	3
I. KISH Plans de sondage à usages multiples	19
P. LAVALLÉE et M.A. HIDIROGLOU Sur la stratification de populations asymétriques	35
D. KASPRZYK Domaines de recherche relatifs à la SIPP (enquête sur le revenu et la participation aux programmes)	47
D. SCHNELL, W.J. KENNEDY, G. SULLIVAN, H.J. PARK, et W.A. FULLER Logiciel d'ordinateur personnel pour l'estimation de la variance dans des enquêtes complexes	63
Section Spéciale - Erreur de couverture dans le recensement	
G. DIFFENDAL Test des opérations de redressement de 1986 dans le Central Los Angeles County ..	75
N. SCHENKER Traitement des données manquantes dans l'estimation de la couverture: le test des opérations de redressement de 1986	93
H. HOGAN et K. WOLTER Mesure de l'erreur dans une enquête post-censitaire	105
P. P. BIEMER Modélisation de l'erreur d'appariement et son effet sur les estimations de l'erreur d'observation du recensement	125

TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada

La revue Techniques d'enquête est répertoriée dans The Survey Statistician et Statistical Theory and Methods Abstracts. On peut en trouver les références dans Current Index to Statistics.

COMITÉ DE DIRECTION

Président

G.J. Brackstone

Membres

B.N. Chinnappa

G.J.C. Hole

C. Patrick

F. Mayda (Directeur de la production)

COMITÉ DE RÉDACTION

Rédacteur en chef

M.P. Singh, *Statistique Canada*

Rédacteurs associés

K.G. Basavarajappa, *Statistique Canada*

D.R. Bellhouse, *U. of Western Ontario*

L. Biggert, *Université de Florence*

D. Binder, *Statistique Canada*

E.B. Dagum, *Statistique Canada*

W.A. Fuller, *Iowa State University*

J.F. Gentleman, *Statistique Canada*

M. Gonzalez, *U.S. Office of*

Management and Budget

D. Holt, *University of Southampton*

K.M. Wolter, *U.S. Bureau of the Census*

V. Tremblay, *Statplus, Montréal*

F.J. Scheuren, *U.S. Internal Revenue Service*

C.E. Sarnadal, *Université de Montréal*

I. Sande, *Statistique Canada*

D.B. Rubin, *Harvard University*

J.N.K. Rao, *Carleton University*

W.M. Podehl, *Statistique Canada*

M.N. Murthy, *Applied Statistics Centre, India*

G. Kalton, *University of Michigan*

POLITIQUE DE RÉDACTION

La revue Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

La revue Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à en faire parvenir le texte au rédacteur en chef, M. M.P. Singh, Division des méthodes d'enquêtes sociale, Statistique Canada, 4^e étage, Édifice Jean-Talton, Tunney's Pasture, Ottawa (Ontario), Canada K1A 0T6. Prière d'envoyer deux exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

Abonnement

Le prix de la revue Techniques d'enquête (catalogue n° 12-001) est de 20,00\$, par année au Canada, et de 23,00\$ par année à l'étranger (paiement en dollars canadiens ou l'équivalent). Prière de faire parvenir votre demande d'abonnement à: Section des ventes des publications, Statistique Canada, Ottawa (Ontario), Canada K1A 0T6. Un prix réduit, soit 10,00\$ (E.-U.) (14,00\$ Can.) est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête et la Société Statistique du Canada. Veuillez envoyer votre demande d'abonnement directement à l'organisation.

TECHNIQUES D'ENQUÊTE

UNE REVUE DE STATISTIQUE CANADA

JUIN 1988

Publication autorisée par
le ministre des Approvisionnements
et Services Canada
©Ministre des Approvisionnements
et Services Canada 1988

Le lecteur peut reproduire sans autorisation des
extraits de cette publication à des fins d'utilisation
personnelle à condition d'indiquer la source en
entier. Toutefois, la reproduction de cette publication
en tout ou en partie à des fins commerciales ou de
redistribution nécessite l'obtention au préalable
d'une autorisation écrite des Services d'édition,
Agent de droit d'auteur, Centre d'édition du gouvernement
du Canada, Ottawa, Canada K1A 0S9.

Septembre 1988

Prix: Canada, \$20.00 par année
Autres pays, \$23.00 par année

Païement en dollars canadiens ou l'équivalent

Catalogue 12-001, vol. 14, n° 1

ISSN 0714-0045

Ottawa

VOLUME 14, NUMÉRO 1
JUN 1988

UNE REVUE
DE
STATISTIQUE CANADA

TECHNIQUES D'Échantillonnage



Statistics Canada Statistique Canada

SURVEY METHODOLOGY

A JOURNAL
OF
STATISTICS CANADA



VOLUME 14, NUMBER 2
DECEMBER 1988

Canada

RENEWAL REMINDER

1989 SUBSCRIPTION

SURVEY METHODOLOGY

A reduced price of U.S. \$16.00, CAN \$20.00 per year is available to members of the American Statistical Association, the International Association of Survey Statisticians and the Statistical Society of Canada. When you renew your membership in your association, please renew your subscription to **Survey Methodology** as well.

Regular subscriptions, in Canadian funds, are \$30.00 in Canada, \$35 for other countries. Please use this form only if you are not subscribing through one of the above associations.

MAIL TO:

Publication Sales

Statistics Canada

Ottawa, Canada K1A 0T6

Name _____

Address _____

City _____

Province _____ Code _____

☐ Enter my subscription to Survey Methodology
(Catalogue 12-001)

☐ Payment enclosed

☐ Bill me later

☐ Mastercard

☐ Visa

Account

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Expiry date

--	--	--	--	--	--

Signature _____

RAPPEL DE RENOUVELLEMENT ABONNEMENT 1989 TECHNIQUES D'ENQUÊTE

Un prix réduit, soit \$16.00 (É.-U.), \$20.00 (CAN) est offert aux membres de l'American Statistical Association, l'Association Internationale des Statisticiens d'Enquêtes et la Société Statistique du Canada. Veuillez renouveler votre abonnement pour **Techniques d'enquête** quand vous renouvelerez votre cotisation.

Le prix régulier en dollars canadiens est de \$30.00 par année au Canada et de \$35.00 par année à l'étranger. Veuillez utiliser cette formule seulement si vous ne vous abonnez pas par votre association

POSTEZ A:

Vente des publications
Statistique Canada
Ottawa, Canada K1A 0T6

Nom _____

Adresse _____

Ville _____

Province _____ Code _____

☐ Veuillez m'abonner pour un an à Techniques d'enquête
(N° au catalogue 12-001)

☐ Paiement inclus

☐ Facturez-moi plus tard

☐ Mastercard

☐ Visa

N° de compte

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Date d'expiration

--	--	--	--	--	--

Signature _____

SURVEY METHODOLOGY

A JOURNAL OF STATISTICS CANADA

DECEMBER 1988

Published under the authority of
the Minister of Regional Industrial Expansion and
the Minister of State for Science and Technology

©Minister of Supply
and Services Canada 1988

Extracts from this publication may be reproduced
for individual use without permission provided the
source is fully acknowledged. However, reproduction
of this publication in whole or in part for purposes of
resale or redistribution requires written permission from
the Programs and Publishing Products Group, Acting
Permissions Officer, Crown Copyright Administration,
Canadian Government Publishing Centre,
Ottawa, Canada K1A 0S9

March 1989

Price: Canada, \$30.00 a year
Other Countries, \$35.00 a year

Payment to be made in Canadian funds or equivalent

Catalogue 12-001, Vol. 14, No. 2

ISSN 0714-0045

Ottawa

SURVEY METHODOLOGY

A Journal of Statistics Canada

The Survey Methodology Journal is abstracted in The Survey Statistician and Statistical Theory and Methods Abstracts and is referenced in the Current Index to Statistics.

MANAGEMENT BOARD

Chairman	G.J. Brackstone	
Members	B.N. Chinnappa	R. Platek
	G.J.C. Hole	D. Roy
	C. Patrick	M.P. Singh
	F. Mayda (Production Manager)	

EDITORIAL BOARD

Editor M.P. Singh, *Statistics Canada*

Associate Editors

K.G. Basavarajappa, <i>Statistics Canada</i>	G. Kalton, <i>University of Michigan</i>
D.R. Bellhouse, <i>U. of Western Ontario</i>	M.N. Murthy, <i>Applied Statistics Centre, India</i>
L. Biggeri, <i>University of Florence</i>	W.M. Podehl, <i>Statistics Canada</i>
D. Binder, <i>Statistics Canada</i>	J.N.K. Rao, <i>Carleton University</i>
E.B. Dagum, <i>Statistics Canada</i>	D.B. Rubin, <i>Harvard University</i>
W.A. Fuller, <i>Iowa State University</i>	I. Sande, <i>Statistics Canada</i>
J.F. Gentleman, <i>Statistics Canada</i>	C.E. Särndal, <i>University of Montreal</i>
M. Gonzalez, <i>U.S. Office of Management and Budget</i>	F.J. Scheuren, <i>U.S. Internal Revenue Service</i>
D. Holt, <i>University of Southampton</i>	V. Tremblay, <i>Statplus, Montreal</i>
	K.M. Wolter, <i>U.S. Bureau of the Census</i>

Assistant Editors

J. Armstrong, J. Gambino and J.-L. Tambay, *Statistics Canada*

EDITORIAL POLICY

The Survey Methodology Journal publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

The Survey Methodology Journal is published twice a year. Authors are invited to submit their manuscripts in either of the two official languages, English or French to the Editor, Dr. M.P. Singh, Social Survey Methods Division, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Two nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of the Survey Methodology Journal (Catalogue No. 12-001) is \$30.00 per year in Canada, \$35.00 per year for other countries (payment to be made in Canadian funds or equivalent). Subscription order should be sent to: Publication Sales, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6. A reduced price of US \$16.00 (\$20.00 Can.) is available to members of the American Statistical Association, the International Association of Survey Statisticians, and the Statistical Society of Canada. Please subscribe through your organization.

SURVEY METHODOLOGY

A Journal of Statistics Canada
Volume 14, Number 2, December 1988

CONTENTS

In This Issue	135
Special Section – Census Coverage Error	
R.D. BURGESS Evaluation of Reverse Record Check Estimates of Undercoverage in the Canadian Census of Population	137
A. ROMANIUC A Demographic Approach to the Evaluation of the 1986 Census and the Estimates of Canada's Population	157
C.Y. CHOI, D.G. STEEL, and T.J. SKINNER Adjusting the 1986 Australian Census Count for Under-Enumeration	173
N. CRESSIE When Are Census Counts Improved by Adjustment?	191
D.B. RUBIN, J.L. SCHAFER, and N. SCHENKER Imputation Strategies for Missing Values in Post-Enumeration Surveys	209
D.J. FEIN and K.K. WEST The Sources of Census Undercount: Findings from the 1986 Los Angeles Test Census	223
M.H. MULRY and B.D. SPENCER Total Error in the Dual System Estimator: The 1986 Census of Central Los Angeles County	241
A.M. ZASLAVSKY Representing Local Area Adjustments by Reweighting of Households	265

CONTENTS – Continued

Software Developments

J. LORIGNY

QUID, A General Automatic Coding Method 289

M.J. WENZOWSKI

ACTR: A Generalized Automated Coding System 299

W. MUDRYK

Quality Control Processing System for Survey Operations..... 309

Y. DeGUIRE

Postal Address Analysis 317

D.N. EMERY

A Brief Note on SQL 327

G. NATHAN

A Bibliography on Randomized Response: 1965-1987 331

Corrigendum..... 347

Acknowledgements 349

In This Issue

Eight papers in this issue deal with **Census Coverage Error**. These papers, together with the four papers on this topic that appeared in the June 1988 issue, provide the reader a good overview of some of the latest methods available for dealing with census coverage error. A great deal of attention has recently been directed at this problem by both policy makers and statisticians. In many countries, studies are carried out during or following each census to measure coverage error. In Canada, the Reverse Record Check (RRC) is the most important study undertaken to measure undercoverage. A Post-Enumeration Survey (PES) is conducted in the United States and Australia.

The papers by Burgess and Romaniuc deal with coverage problems in the Canadian Census of Population. Burgess describes the RRC methodology, and considers some of its limitations that lead to errors in estimates of undercoverage. Romaniuc, on the other hand, takes a demographic approach to the study of the accuracy of the census. The results obtained in this way are contrasted with those based on the RRC. In addition, Romaniuc looks at the quality of data for components of change (births, deaths, migration) used in the demographic approach.

Choi, Steel and Skinner's paper deals with the 1986 Australian PES. Like Romaniuc, the authors consider demographic estimates of under-enumeration. Based on their analysis, the authors conclude that PES-based adjustments should continue to be used in the 1991 Census, but emphasize that investigation of bias problems should continue.

Cressie uses a model for undercount errors to investigate the adjustment of census counts. He considers synthetic estimation, Bayes and empirical Bayes approaches, and uses risk to compare estimators. A "usual empirical Bayes" estimator is found to have the smallest risk. Cressie notes that the results depend on the assumption that a sufficiently large number of households are chosen in the PES.

The paper by Rubin, Schafer and Schenker on imputation for missing values in a PES also has a Bayesian flavour. The authors review the imputation methods discussed by Schenker in the previous issue of **Survey Methodology**. They propose two model-based methods, and conclude that the method that does not ignore the missing data mechanism is preferable. The authors caution that, although their approach looks promising, more work is needed.

Fein and West present a systematic classification of the causes of undercount and conclude that partial household omission is the biggest contributor to the undercount. Methodological analysis of total error in the dual system estimator (an estimator that was examined by authors in the June 1988 issue) is discussed by Mulry and Spencer. Using a Bayesian approach, the authors combine the error components to obtain a final interval estimate of net undercount rate.

Zaslavsky deals with the undercount problem by using block-level undercount estimates to reweight households in the block. An advantage of this approach is that the "character" of each block is preserved. The details of the method are interesting and will look familiar to readers acquainted with raking methods.

The development of new computer systems designed to process large amounts of information is a topic of increasing interest to survey statisticians. Five of the papers in this issue describe **Software Development** related to survey methodology.

Automated coding systems developed by central statistical agencies are described in two papers. Lorigny deals with the QUID system used at the Institut National de la Statistique et des Études Économiques. Wenzowski's paper is a guide to the ACTR system, developed at Statistics Canada. Both QUID and ACTR are designed to handle any type of classification system efficiently.

Readers will be interested in comparing the approaches taken in the two systems. Some performance data are also given.

Mudryk describes a computer system for quality control currently used as part of Statistics Canada's overall quality assurance program. The objectives of the system are both to exercise error prevention in survey processing operations and to reduce inspection levels progressively as the quality of processing improves and stabilizes.

Deguire describes a system, designed to analyze the syntax of postal addresses, currently under development at Statistics Canada. The software produces address search keys consisting of standardized address components that can be used during computerized matching operations such as those involved in the construction of a national Address Register.

Emery describes SQL (Structured Query Language), the most popular query language associated with relational database management systems. The strengths and weaknesses of the language are highlighted.

In the final paper in this issue, Nathan provides a comprehensive list of over 250 books, theses and papers dealing with randomized response. A subject classification is also included.

The Editor

Evaluation of Reverse Record Check Estimates of Undercoverage in the Canadian Census of Population

R.D. BURGESS¹

ABSTRACT

Estimates of undercoverage in the Canadian Census of Population have been produced for each Census since 1961, using a Reverse Record Check method. The reliability of the estimates is important to how they are used to assess the quality of the Census data and to identify significant causes of coverage error. It is also critical to the development of methods and procedures to improve coverage for future Censuses. The purpose of this paper is to identify potential sources of error in the Reverse Record Check, which should be understood and addressed, where possible, in using this method to estimate coverage error.

KEY WORDS: Matching; Mobility; Nonresponse bias; Response error; Reverse record check; Sampling error; Tracing.

1. INTRODUCTION

The Census of Canada is conducted every five years; the most recent was in 1986. Starting with the 1971 Census, the main data collection methodology has been self-enumeration: less than 4% of the population are enumerated using the canvasser method. In geographic areas where self-enumeration is used, each dwelling is listed and a questionnaire dropped off by an enumerator just prior to Census Day (June 3 in 1981 and 1986). In larger urban areas the respondent household is asked to return the completed questionnaire by mail to the local supervisor of the enumeration. In rural areas and smaller urban areas the questionnaires are picked up by the enumerator.

The enumerator is to perform basic checks of coverage and response quality for his/her assignment and follow up on missing and incomplete questionnaires. Supervisory checks and quality control of the enumerator's work are also carried out. However, there is no independent and rigorous check of the listing of dwellings. Further, there is only limited opportunity to verify the number of persons listed on the questionnaire by the respondent household.

Not unexpectedly there are overcoverage and undercoverage errors in the Census. Such errors are important because of the various uses of Census data; representation in the Parliament of Canada is determined using Census population counts; various federal-provincial government financial agreements incorporate formulae that have population count or distribution as a factor (Statistics Canada 1983b). In turn the quality of estimates of coverage error is an important issue: for the use of Census data; in considering adjustment of population and dwelling counts to compensate for the coverage error; and in attempting to improve coverage quality for future Censuses by identifying significant causes or areas of coverage error.

Since 1961, Statistics Canada has produced and published an estimate of undercoverage for each Census of Population. The method used to produce these estimates has been a Reverse Record Check (RRC) study which involves five general activities or stages:

¹ R.D. Burgess, Social Survey Methods Division, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, K1A 0T6.

- (i) frame preparation – identification of a set of nonoverlapping lists that together are to cover the total population that should be enumerated in the Census;
- (ii) sample design and selection – selection of a random sample of persons from the lists;
- (iii) tracing – determination of the address of usual place of residence on Census Day for each selected person (or verification that he/she died or emigrated prior to the Census);
- (iv) searching – review of Census returns to determine whether the selected person had been enumerated or missed in the Census; and
- (v) weighting and estimation – weighting up of sample results to produce an estimate of the number of persons missed in the Census.

A more detailed description of this methodology can be found in Gosselin 1976 or Statistics Canada 1984.

Other methodologies – post Census re-enumeration, demographic analysis and administrative record checks – could also be used to estimate Census undercoverage. In the Canadian context, however, each of these methodologies would likely produce results less reliable than those of the RRC. Re-enumeration studies show a tendency to miss the same households or persons as the Census itself. Demographic methods are model-based and suffer from a lack of reliable emigration estimates, measure only change in net coverage between censuses, do not identify individual cases and causes of coverage error, and are weakened sub-nationally by error in internal migration estimates. Administrative record checks are limited by the absence of a national administrative system that either has more complete coverage than the Census or has coverage errors independent of Census coverage error – a condition that would allow an incomplete administrative file to be used. Even if such a complete system existed, its use would be another version of a reverse record check, unless it were completely up to date in coverage and addresses, as of Census Day.

For these reasons the reverse record check has been the preferred methodology in Canada, though demographic analysis methods have been used for corroborative analysis. However, the RRC itself has deficiencies. The purpose of this paper is to describe some of the sources of error or limitations in the RRC method, in the context of the Canadian Census of Population. In Section 2 aspects of the survey methodology of the RRC that can lead to error in the final results are reviewed. The results of some analysis of RRC estimates, in conjunction with data from other sources, have raised unresolved problems related to the use of RRC results in population estimation. These results are presented in Section 3. Some concluding remarks are given in Section 4.

2. LIMITATIONS OF THE REVERSE RECORD CHECK METHODOLOGY

A limitation, in the context of this paper, is anything that restricts the applicability of the Reverse Record Check estimates or the confidence with which they can be used. Limitations can arise because of: differences between what is conceptually required by users and what the RRC attempts to measure; shortfalls in the design of the Reverse Record Check in attempting to meet its objectives; or sampling, response and other errors. Some of these limitations might be eliminated or reduced through modification of specific aspects of the Reverse Record Check. Others will persist or, by their nature, cannot be addressed.

2.1 Applicability of Reverse Record Check Estimates

The objective of the Reverse Record Check is to provide estimates, for each of the ten provinces, of undercoverage in the Census of Population. Net coverage error is not estimated and the Yukon and Northwest Territories are excluded from the study.

The RRC estimates the proportion of the population missed in the Census – *i.e.*, the proportion of the population that was not enumerated but should have been. Overcoverage (persons enumerated more than once, and persons enumerated who should not have been or were fictitious) is not estimated by the RRC. Thus net coverage error, undercoverage minus overcoverage, is not estimated by this vehicle. Even if the amount is small, the potential importance of overcoverage lies in its size and distribution relative to undercoverage. For example, overcoverage of 0.2%, one tenth the level of undercoverage in 1976 and 1981, would be very important if the rate for a particular province is as high as 0.5%.

The two Canadian territories have not been included in the RRC because the size of their populations is small but they have exceptionally high rates of intercensal in and out migration. In terms of sampling error, to produce reliable estimates for the territories, a proportionally large sample of the territorial population would have to be selected – of the order of a 5% sample or 3,750 persons. The territories have in and out intercensal migration rates of a third or more. Therefore, 1,250 of the 3,750 persons (on average) in the minimum sample should be intercensal in-migrants, assuming a proportional sample is required. The RRC uses lists for which the address of residence for the majority of persons was obtained five years earlier and in-migrants to the territories can only be identified during the conduct of the study. This in itself is not a problem. However, the RRC uses only a 0.15% sample. The in-migrants to the territories, therefore, would be expected to be sampled at this latter rate and not at the required 5% rate. This would result in a sample of in-migrants to the territories of only 30 persons. Thus, within the current framework of the RRC, and without prohibitive additional expense, it is not possible to select a meaningful sample to represent that third or more of the territorial population who are intercensal in-migrants.

2.2 The Reverse Record Check Methodology

Each of the five stages of the Reverse Record Check is a known or potential source of error.

2.2.1 Frame

The sample for the RRC is selected from four lists or frames:

- (i) Census: persons enumerated in the previous Census – for example, the 1981 Census was used for the 1986 Reverse Record Check;
- (ii) Birth: intercensal births, obtained from vital statistics records;
- (iii) Immigrant: intercensal immigrants, obtained from records of Employment and Immigration Canada; and
- (iv) Missed: persons missed in the previous Census – which is available as a sample only from the previous Reverse Record Check (no complete list exists for this group).

These lists are intended to include or represent, without duplication of individuals on or between lists, all persons who should be enumerated (in one of the ten provinces) in the current Census.

Some people, however, are not represented on these lists. Included among these are: (a) intercensal and never enumerated illegal aliens; (b) certain classifications of refugee; (c) certain

Canadians "abroad" at the time of the previous Census who returned prior to the current Census; (d) persons who move from the territories to one of the provinces in the intercensal period; and (e) persons not enumerated in any Census covered by the application of the RRC, but who were usual residents of Canada prior to 1961.

It is assumed, without direct evidence, that the number of persons in category (e) has become small enough to be irrelevant. For the 1981 Census the size of category (d) was estimated to be of the order of 18,000 persons. Most of these persons were usual residents of the territories at the time of the previous (1976) Census. There were probably also a few of what would be Birth frame and Immigrant frame persons among the 18,000.

Category (c) includes some Canadians working, studying or travelling abroad who did not maintain a usual place of residence in Canada during their absence and may also include children born outside Canada to parents in this category. It does not include persons in the Canadian military, in External Affairs or other government service (and their families) living abroad. They are included in the Census frame and the Missed frame. For the 1981 Census, the size of this returning "abroad" group was estimated to be approximately 67,000 persons.

Refugee applicants and illegal aliens in Canada are to be enumerated in the Census, assuming they do not have a usual place of residence outside of Canada, and are not holders of work or student visas. For the 1981 and 1986 RRC studies, persons applying from abroad and entering Canada as refugees were included in the Immigrant frame. Persons applying within Canada were included in the Immigrant frame only if they had been granted refugee status. As of April 1985, there were 12,500 applications from within Canada under consideration PLAUT 1985. The number of illegal aliens in Canada is not known or reliably estimated. Some illegal aliens may be represented in the Census frame or even the Missed frame. Amnesty programmes in the 1970's and 80's will have resulted in some illegal aliens being entered in the Immigrant frame.

Under the current RRC methodology the exclusions to the frames are important to the extent that such persons are not counted in the current Census. Since the Immigrant frame tends to have a high undercoverage rate (8.5% compared to 2.0% overall in 1981), it is not unreasonable to expect a high undercoverage rate for the refugee status claimants. It is possible that the majority of illegal aliens were not counted in the Census. These elements of undercoverage could be significant relative to the estimated number of persons missed (approximately 500,000 in 1981). The refugee status claimants and the illegal aliens may have been clustered in a few urban centres within only certain provinces. This would increase the impact of such exclusions on the reliability of estimates.

The lists can also be expected to include some amount of overcoverage; *e.g.*, persons enumerated in the previous Census who should not have been or who were enumerated more than once, fictitious persons and processing errors. Some overcoverage is detected during the course of the RRC operations. In estimating undercoverage, however, the effect of overcoverage in the frames would be consequential only if it approaches or exceeds the undercoverage in the Census in size.

2.2.2 Sample Size and Design

Error due to sampling is a major limitation of the RRC results. While the potential size of this error is dependent upon sample size and design, the sample size is the more important element. It, along with the available lists, limits the design options.

The basic 1981 and 1976 RRC undercoverage estimates for provinces and their corresponding estimates of standard error are presented in Table 1. The coefficients of variation (standard error divided by estimated undercoverage) varied from 4.5% at the Canada (10 provinces)

Table 1

Estimated Population Undercoverage in the 1981 and 1976 Census,
by Province, showing Provinces with Significant Differences in
Population Undercoverage (with 95% confidence)

Province	Population Undercoverage		Province with a Significantly Different Undercoverage Rate
	Rate	S.E.	
	(%)	(%)	
1981 Census			
Canada (10 Provinces)	2.01	0.09	
1. Newfoundland	1.74	0.45	10
2. Prince Edward Island	1.17	0.54	9 and 10
3. Nova Scotia	1.05	0.34	5, 6, 9 and 10
4. New Brunswick	1.81	0.30	10
5. Québec	1.91	0.21	3, 7, 8 and 10
6. Ontario	1.94	0.14	3, 7, 8 and 10
7. Manitoba	0.98	0.35	5, 6, 9 and 10
8. Saskatchewan	0.99	0.37	5, 6, 9 and 10
9. Alberta	2.54	0.36	2, 3, 7 and 8
10. British Columbia	3.16	0.33	all but 9
1976 Census			
Canada (10 Provinces)	2.04	0.10	
1. Newfoundland	1.10	0.39	5 and 10
2. Prince Edward Island	0.38	0.25	4, 5, 6, 8, 9 and 10
3. Nova Scotia	0.86	0.34	4, 5 and 10
4. New Brunswick	2.16	0.37	2, 3, 7 and 10
5. Québec	2.95	0.25	1, 2, 3, 6, 7, 8 and 9
6. Ontario	1.52	0.17	1, 2, 5, 7 and 10
7. Manitoba	1.07	0.33	4, 5 and 10
8. Saskatchewan	1.33	0.34	2, 5 and 10
9. Alberta	1.49	0.26	2, 5 and 10
10. British Columbia	3.13	0.31	all but 5

level, up to 13.6% at the regional (Atlantic, Québec, Ontario, Prairie and British Columbia) level and up to 46% at the provincial level. Sub-provincial coefficients of variation were typically higher. For an Electoral District of average size (86,323 persons in 1981) with an estimated 2% undercoverage, the coefficient of variation would be approximately 50%. For smaller geographic areas and small population groups the coefficient of variation could be much higher.

The sampling error, of course, has an effect on attempts to differentiate among provincial, and among other undercoverage rates. In turn this affects attempts to identify specific causes or areas of undercoverage, and undermines the validity of adjusting for coverage error as a means to improve Census counts. Those provinces with a significantly different undercoverage rate are also shown in Table 1. The undercoverage rates for the provinces appear to fall into six groupings, for 1981, based on both rate of undercoverage and provinces with which the rate is significantly different. For 1976, with eight groups, there was less similarity between provinces. No group in either Census, however, can be shown to be completely different from all others, and may not be.

This general situation is not dissimilar to that for applications of the Reverse Record Check for the 1966 and 1971 Censuses. From 1966 onward only the province of British Columbia has had an undercoverage rate significantly above the Canada level. The variation from Census to Census for most provinces, in large part, could be due to sampling error. Why it is not for British Columbia is a major concern for both the Reverse Record Check and the Census.

The need to use a sample of "missed" persons from the previous RRC also places a limitation on the design and sample size. There is no direct control of the size of this segment of the sample. Any limitations of the previous Reverse Record Check, to the degree that these were reflected in the estimate of "missed" persons, will be passed. (See Sections 2.2.4, 2.2.5 and 3).

2.2.3 Tracing

Given the nature of the lists or frames used for sample selection, addresses and other information may be up to five years out of date. Attempts are made to update addresses prior to Census Day using administrative files. (This was first carried out extensively for the 1986 RRC.) After Census Day, the Census questionnaire corresponding to the original address, or the update if available, is searched as a first attempt to determine whether the selected person was enumerated in the Census. Every selected person not found enumerated in the first search must be traced. The selected person, or a reliable source, must be contacted either to obtain an updated or confirmed address, or to determine the selected person's status, *i.e.*, as deceased, emigrated, abroad.

Despite extensive tracing activities, not all selected persons can be traced. This may result in a form of nonresponse bias. In the 1981 RRC 3.4% of all selected persons were not traced. With overall undercoverage in the Census estimated to be 2.0% this "not traced" rate represents an important uncertainty in the RRC estimates.

A weight adjustment is carried out to account for these "not traced" cases. The effect of the weight adjustment for the 1981 Census was to impute an undercoverage rate of 3.27% for the "not traced" cases from the Census and Missed frames (jointly), 1.46% for the Birth frame and 11.94% for the Immigrant frame. Overall, the proportion of "not traced" weights "imputed" by the weight adjustment to "missed" was 1.6 times the initial (weighted) proportion represented by the "missed" cases among all traced selected persons. This suggests a relationship between "not traced" and "missed". It is not known, of course, if the 1.6 rate was too high, too low or correct. To the extent that it is not correct, there may be some distortion in provincial estimates of undercoverage as well as a bias in overall estimates of undercoverage.

Since the rates of intercensal interprovincial in and out-migration vary from one province to another, there may be some distortion among provincial estimates. This will occur if the proportion of interprovincial movers within weighting groups is not the same among the cases traced and not traced.

Intercensal interprovincial movers (applicable for Census and Missed frames only) have a high undercoverage rate. This rate was estimated to be 6.13% for the 1981 Census, based upon mobility data from the 1981 RRC derived by comparing of the 1976 Census and 1981 Census addresses. The estimated undercoverage rate for intercensal migrants within a province (*i.e.*, between Census Subdivision (CSD) or municipality movers) was 3.83%. For intercensal non-migrant movers (within CSD or municipality) the undercoverage rate was estimated to be 2.83%. Given these rates and the distribution of mobility characteristics, the "imputed" undercoverage rate for the "not traced" cases from the Census and Missed frames put together would be expected to be at least 3.52% rather than the actual 3.27%. That is, given persons not traced almost always have moved. It is, in turn, assumed that these "not traced" cases included

proportionally at least as many migrants, within and between provinces, and had not less than the same undercoverage rates, by mobility status, as traced cases. (The distribution of mobility status of the enumerated population 5 years and older, estimated through the 1981 RRC was approximately: (i) Non-movers – 55%; (ii) Non-migrant Movers – 17%; (iii) Migrants Same Province – 21.7%; (iv) Migrants Different Province – 5%; and (v) Migrants From Outside Canada – 2%.)

Given the tracing methods used, it is not unreasonable to speculate that the proportion of migrants, and thus the undercoverage rate, was much higher for the “not traced” cases. If they were, then there could be a significant downward bias in the estimates of undercoverage. For example, if the “true” undercoverage rate among the cases not traced was close to 5.0%, then the bias in the undercoverage estimate at the Canada (10 provinces) level would exceed the sampling error.

2.2.4 Searching and Classification

After all tracing attempts have been made and any interviews conducted, each selected person is classified to one of six categories:

- (1) enumerated;
- (2) missed;
- (3) deceased;
- (4) emigrated or abroad;
- (5) overcoverage in a list or frame; and
- (6) not traced.

As outlined above, to determine whether a selected person has been enumerated or missed the Census questionnaire corresponding to the selected person’s address must be searched. For the search to result in the correct classification of the selected person, it is necessary that the address being searched be the correct address, and that the selected person be correctly identified on the Census questionnaire and in RRC documentation; *i.e.*, that there be no response error or nonresponse for the relevant items.

If the selected person is correctly identified (complete name, correct age and sex, *etc.*) and there are no processing errors, then no selected person who was missed in the Census will be classified as “enumerated”. The converse is not true. If a selected person has been enumerated in the Census at some address other than that which is obtained from the list of selection, some other administrative source or a directory, then to be classified as “enumerated” that address must be provided by the selected person or some other contact. If the selected person does not or can not provide that address (for example, recall error or can not remember), then he or she will be classified as “missed” or “not traced”. Generally, when the selected person (or a parent, spouse or other reliable source) gives an address, or set of addresses, where he/she should have been or may have been enumerated, this address information is accepted as correct. Selected persons will be classified as “enumerated” or “missed” based on this address information. It is not known how accurate such address information actually is for persons classified as “missed”.

On the other hand there may be a higher probability of classifying a person missed as “not traced” than a person enumerated in the Census. Before a person can be classified as missed he/she (or a reliable source) must be interviewed to confirm the address and to obtain possible alternative addresses and certain Census data for him/her and the household. This procedure will eliminate some classification error. At the same time, if the information about a person missed is doubted this can only be resolved through the contact with him/her (or a parent,

spouse, etc.). If the doubt is not resolved the case will be classified as "not traced". Conclusive information is not always necessary for a person who was enumerated. With exhaustive searching it may be possible to transform a selected person, who was enumerated, from "not traced" to "enumerated", even if the address obtained is incomplete or incorrect. Such searching is much less likely to alter the outcome for persons missed in the Census.

The selected person is not always adequately identified. In accepting a selected person as matched; *i.e.*, found enumerated on a Census questionnaire – name is not always identical on the Census and RRC documents. Sometimes only the first person listed on a Census questionnaire has a complete name and in a few cases no names are given. If the identity of the selected person cannot be determined from the list or frame, then the case will be classified as "not traced" at the outset. Included among these will be persons "assigned" for absent households and refusals in the previous Census. Date of birth and other data are not always present, complete or found identical in matching. For the majority of cases the quality of matching is unquestioned, but a minority of cases raise doubts. Doubtful cases accepted as matched potentially are misclassified as "enumerated". Those rejected as matched potentially are misclassified as "missed", though most will be classified as "not traced". Different rules for acceptance/rejection as matched, of course, may yield different estimates of undercoverage.

Some overcoverage in the frames can be detected. This will include: some foreign residents enumerated in the previous Census; persons "created" by processing error in the previous Census; immigrants who have not yet resided in Canada; births in Canada to non-resident parents; and fictitious or out of scope "persons" listed on the questionnaire from the previous Census. In 1981 these cases represented less than 0.1% of selected persons.

Overcoverage in the form of duplication in a frame will not be detected. Fictitious selected persons may go undetected and be classified among the "not traced" cases.

The final classifications of the selected persons from the 1981 RRC are presented in Table 2 (from Burgess 1986).

2.2.5 Weighting and Estimation

At the time of sample selection, a basic weight equal to the inverse of the sampling fraction is assigned to each selected person record. Two types of weight adjustment are made to this basic weight – one to account for "not traced" cases, the other to account for deviations in

Table 2
1981 RRC Final Classification of Selected Persons

Final Classification	Frame									
	Census		Birth		Immigrant		Missed		Total	
	Cases	%	Cases	%	Cases	%	Cases	%	Cases	%
Traced	29,761	97.1	3,211	92.3	1,392	96.1	807	96.1	35,171	96.6
Enumerated	27,541	89.8	3,096	89.0	1,113	76.8	696	82.9	32,446	89.1
Deceased	1,056	3.5	33	0.9	5	0.3	26	3.1	1,120	3.1
Emigrated/Abroad	299	1.0	34	1.0	111	7.7	24	2.8	468	1.3
Missed	865	2.8	48	1.4	163	11.3	61	7.3	1,137	3.1
Not Traced (incl. Overcoverage)	895	2.9	267	7.7	57	3.9	33	3.9	1,252	3.4
TOTAL	30,656	100.0	3,478	100.0	1,449	100.0	840	100.0	36,423	100.0

the representativeness of the sample, after elimination of “not traced” cases, relative to the lists of selection.

A “not traced” case represents a person enumerated or missed in the Census, a deceased person, an emigrant, a person abroad or overcoverage. The weights of the “not traced” cases, therefore, are redistributed among the “traced” cases. The adjustment is carried out within groups defined by various demographic and geographic characteristics, and frame.

The weight adjustment for the “not traced” cases is carried out in two stages. First, an adjustment is made for those cases for which no tracing was undertaken because there was inadequate information for matching and tracing. These cases are weighted into all other selected persons. Second, an adjustment is made for all other “not traced” cases. These are weighted into specific groups of the remaining selected persons. How the “not traced” adjustment is carried out is restricted by the information available on the “not traced” selected persons. Ideally, how a selected person was traced and whether he/she had moved and how far, as well as demographic characteristics, should be taken into consideration in defining weighting groups. To date only demographic characteristics and minimal mobility data have been used in the weight adjustment. (Persons selected in the Census frame who have not moved in the intercensal period and who were classified as “enumerated” are excluded from this weight adjustment.) By their nature it is difficult to categorize most “not traced” cases beyond the fact that they were not found enumerated at the address given on the list of selection.

For the second type of adjustment, totals for relevant sub-groups of the population are obtained from each frame (except for the Missed frame for which only a sample is available). Using these “known totals”, an adjustment to the RRC weights is made within the corresponding subgroups of the sample. This is done to reduce the error in the estimates by ensuring that totals from the sample, for basic population characteristics for which undercoverage rates are published, correspond to the totals in the frames.

Neither adjustment deals at all with the various exclusions to the lists used for sample selection. In the calculation of any proportion of persons missed in the Census the published Census count of enumerated persons is used in the denominator in order to minimize sampling error. (The covariance of the estimate of “enumerated” persons and the estimate of “missed” persons tends to be negative.) Since the RRC does not represent all elements of the true population, the effect of using the Census count is to assume that the undercoverage rate for the exclusions is zero.

The estimator, which takes the general form defined as:

Estimated proportion of persons missed

$$= \frac{\text{Estimated no. of missed persons}}{\text{no. of persons counted in the Census} + \text{Estimated no. of missed persons}}$$

is discussed further in Appendix 2.

2.3 Reducing Potential for Error and Methodological Limitations

Experimental work and evaluation of methods in the RRC may make it possible to eliminate or reduce the impact of some sources of error or limitations.

Overcoverage might be estimated by means of an independent study. Such a study is being conducted, on an experimental basis, for the 1986 Census. However, the cost to produce estimates of adequate quality at the province level may be very high.

The production of estimates for the Yukon and Northwest Territories requires a set of lists other than those used for the RRC. Such a set would have to be current and have no significant duplication that could not be removed or estimated. With such a set of lists, the basic

RRC methods could be applied. Some experimental work in this regard has been done and more is planned.

The lists used for the RRC could be augmented to eliminate some of the exclusions, for example, refugee status claimants and migrants from the territories to the provinces. These people, however, will be difficult to trace. Sampling these groups may do little more than change the nature of the problem.

A sample of "abroad" persons could be obtained by using the previous Reverse Record Check. Such a sample, however, would be very small, would not represent the entire group in question and the selected persons would be difficult to trace.

Other than illegal aliens the "never enumerated" group will become smaller and smaller over time. Intercensal illegal aliens, and other illegal aliens never enumerated in Canada, will remain excluded.

The impact of sampling error can be reduced by increasing the sample size. The question is to what size, at what cost, based upon what criteria? An increase in the RRC sample from its current 36,500 persons to 100,000 should be sufficient to bring the provincial standard error estimates, for the undercoverage rates, down below 0.2%. However, this may not be sufficient for purposes of adjusting the Census counts, depending upon the level and distribution of undercoverage estimates actually obtained. A reduction of the standard error to 0.1% for each province – the level yielded by the 1981 and 1976 RRC studies for the Canada (10 province) level estimate of undercoverage of 2% – would require a sample for Canada of approximately 350,000 persons, assuming the 1981 provincial levels of undercoverage, type of sample design and design effects. To conduct a high quality RRC operation for such a large sample, given the controls and quality checks required, would be much more costly than the mere increase in sample size suggests, and might be operationally unrealizable. Increasing the sample size, of course, would not reduce any bias in the estimates.

Tracing methods are examined before and after each RRC. Major changes were made for 1986 and changes and improvements are being contemplated for 1991. It must be expected, however, that there will again be a non-negligible percentage of "not traced" cases. These cases will continue to be dealt with by weighting or by imputation and weighting.

Evaluative studies can be conducted to assess the quality of matching and of address information provided by respondents or reliable sources. The potential impact of the matching algorithm or criteria can also be assessed to some extent. However, even if such studies identify a problem, solutions may not be readily forthcoming.

Modifications to the weighting procedures can be tested in an attempt to better deal with mobility and other characteristics when adjusting for "not traced" cases (Burgess 1986). Additional information for this purpose might be available from administrative sources. Some minor refinements using existing information can also be made. For example, the adjustment for "not traced" persons contacted, but from whom the necessary Census Day address information could not be obtained, might be different from that for "not traced" persons who potentially may be "deceased", "emigrated" or "abroad".

Adjustments using current Census totals of enumerated persons could be tested as well. For this to reduce any bias associated with "not traced" cases and persons not represented in the RRC sample, however, the basic classification of cases to "missed" must be without bias and there must be no interprovincial distortion of the proportion "missed". These types of modifications to the weighting would not in themselves eliminate bias.

3. ANALYSIS OF REVERSE RECORD CHECK RESULTS

The RRC not only provides estimates of the number of persons missed in the Census, but also independent estimates of the number of persons enumerated in the Census, and the number of intercensal deaths, emigration and persons who have moved abroad but who have not emigrated. These estimates are used in validating RRC estimates. Some of the results of this validation process serve to illustrate limitations discussed in Section 2.

Analysis has also been carried out to correlate geographic variation in undercoverage to variation in the distribution of Census population and household characteristics.

3.1 Independent Estimates

The Reverse Recorded Check estimates of persons enumerated in the Census, of intercensal deaths, and of persons leaving Canada in the intercensal period can be compared to estimates from other appropriately chosen sources – for example, estimates of enumerated persons to Census counts and estimates of deaths to Vital Statistics data. If there are no significant biases in the RRC estimates, then any differences between these estimates will usually be explainable by the corresponding sampling error of the RRC estimate. If there are significant differences, then these might be due to biases in the RRC estimates. The overall quality of these estimates, revealed by the comparisons, likely will be a reflection of the quality of the estimates of “missed” persons.

RRC estimates of emigrants (296,727) and of persons “abroad” (57,909) compared favourably with estimates based upon demographic analysis. The RRC estimate for emigrants, for example, is in the mid range of the five demographic analysis values examined – ranging from 197,000 to 372,000, with a mean value of 266,400. The RRC estimate of deceased persons (846,378) is very close to the value (840,689) published by Statistics Canada 1976 to 1981.

Comparisons of estimates for enumerated persons do indicate some problems. Some of these comparisons are presented in Table 3. For Canada (10 provinces) and for two of the ten provinces, the number of persons enumerated in the Census, as estimated by the RRC, is significantly different from the published Census count. The discrepancy of 209,911 at the aggregate level can be explained in part by exclusions from the lists or frames of the RRC. The discrepancies among provinces is difficult to explain. That in particular makes the discrepancy important. The 209,911 aggregate discrepancy must be considered in the context of the RRC estimate of 497,277 persons missed in the Census; similarly, the discrepancy for British Columbia of 80,304 in the context of an estimated 89,445 persons missed and the discrepancy for Alberta of 86,244 persons in the context of an estimated 58,335 persons missed.

An estimated 67,000 non-immigrants who had been “abroad” at the time of the previous Census arrived in Canada legally, and an estimated 18,000 persons moved from the territories to a province in the intercensal period. Assuming none of these people was missed in the Census, the discrepancy would be reduced to approximately 125,000 persons. This difference would remain at the outer limits of what would be reasonably accepted as due to sampling error only. Further, all of these 85,000 (67,000 + 18,000) persons would have had to have moved to Alberta and British Columbia to reduce the discrepancies for these provinces to within 95% confidence intervals – a clearly unreasonable supposition.

The remainder of the difference (125,000) could be made up of various (potential) errors in the RRC or the Census: (i) sampling error in the RRC estimate of enumerated persons; (ii) an increase in overcoverage in the 1981 Census – compared with the 1976 Census; (iii) RRC exclusion of illegal aliens and refugee claimants enumerated in the Census itself; (iv) underestimation of persons missed in the 1976 Census – these persons make up 1981 Missed

Table 3
Reverse Record Check Estimates of the Number of Persons
Enumerated in the 1981 Census by Province

Province	RRC Estimate of Persons Enumerated	S.E. of RRC Estimate	Census Published ¹ Count	Persons Enumerated RRC-Census	RRC Estimate of Persons Missed
Canada (10 provinces)	24,064,376	62,193	24,274,287	-209,911 ²	497,277
Newfoundland	568,696	8,256	567,681	1,015	10,039
Prince Edward Island	116,012	3,005	122,506	-6,494	1,456
Nova Scotia	837,045	11,185	847,442	-10,397	9,034
New Brunswick	685,332	8,167	696,403	-11,071	12,864
Québec	6,410,662	38,648	6,438,403	-27,736	125,180
Ontario	8,629,374	52,802	8,625,107	4,267	171,010
Manitoba	1,028,162	15,133	1,026,241	1,921	10,203
Saskatchewan	973,450	11,740	968,313	5,137	9,712
Alberta	2,151,480	24,238	2,237,724	-86,244 ²	58,335
British Columbia	2,664,163	19,798	2,744,467	-80,304 ²	89,445

¹ Statistics Canada 1982.

² Greater than 3 standard errors.

frame; and/or (v) over-estimation of persons missed in the 1981 Census. The extent to which each of these sources might have contributed to the difference is not known. The fact that a large part of the difference seems to be associated with British Columbia and Alberta is perhaps in some degree due to under-estimation of intercensal migrants. Migration to these provinces was particularly high between 1976 and 1981 (Statistics Canada 1979; 1983a).

There may also be some bias in the estimates of emigrated, abroad and/or deceased persons. If these are over-estimated for reason other than "not traced" bias, there should also be a tendency to under-estimate the persons missed, since the last address in Canada is sought and used in searching. Persons who emigrated, died or went abroad after Census Day may have been reported as such at the time of tracing, perhaps several months after Census Day. At the same time, the fact that deceased persons do not appear to have been under-estimated despite the exclusions to the RRC frames suggests a lower mortality rate for the exclusions (as is the case for immigrants - see Table 2) than for the entire population and/or over-estimation of this group.

The data in Table 4 show that intercensal migrants were under-estimated for all provinces except Saskatchewan. This may be in part associated with the "not traced" cases. The under-estimation for British Columbia may explain the discrepancy for this province shown in Table 3. On the other hand, the under-estimation for Alberta does not adequately explain the discrepancy for that province and, thus one or more of the factors (i) to (v) noted above must be contributing to this discrepancy.

Under-estimation of migrants may cause a distortion of undercoverage estimates among the provinces; *i.e.*, the large differences shown in Table 4 by province might be indicative of substantial biases in provincial under-enumeration rates. Further, as noted in Section 2.2.3, migrants have higher than average levels of undercoverage. If the enumerated persons within this group are under-estimated, while in general non-migrants are not under-estimated, relative to the Census, then estimates of undercoverage may be too low.

Table 4
Reverse Record Check Estimates of Migrants¹ Enumerated
in the 1981 Census, by Province

Province	Estimate of Migrants			Census Estimate of Inter-Provincial Migration		
	RRC	Census published estimate ²	Difference RRC-Census	In	Out	Out/In
Canada	4,670,311	5,046,500	-376,239	1,124,970	1,122,370	-
Newfoundland	61,499	72,100	-10,601	18,430	38,265	2.08
Prince Edward Island	13,257	20,530	- 7,273	9,945	9,950	1.00
Nova Scotia	125,949	137,865	-11,916	54,455	62,880	1.16
New Brunswick	96,607	109,955	-13,348	41,460	49,965	1.21
Québec	1,092,919	1,145,085	-52,166	61,310	203,035	3.31
Ontario	1,572,504	1,725,225	-152,721	250,570	328,640	1.31
Manitoba	143,391	165,105	-21,714	54,030	97,620	1.81
Saskatchewan	204,937	192,840	12,097	63,395	69,220	1.09
Alberta	669,995	691,970	-21,975	336,830	139,180	0.41
British Columbia	689,253	785,825	-96,622	234,545	123,615	0.53

¹ A migrant is a person who at the time of the previous Census was living outside Canada, in a different province or in a different municipality (or CSD). RRC mobility data used here are those given by the RRC sample person in the Census and not those derived within the RRC (based upon a comparison of addresses).

² Statistics Canada 1983a.

Discrepancies between the RRC estimate of enumerated persons and the Census count have also occurred for earlier Census. The value of the RRC estimate minus the Census count was 289,000 for 1971, and -324,000 for 1976. For both of these Censuses, the RRC estimates of persons deceased and emigrated/abroad were consistent with other sources. The large change from 1971 to 1976, coincident with the large negative values for two consecutive Censuses, cannot emanate from a single source. Changes in the size of overcoverage, larger than the size of the discrepancies, would be required between Censuses. This by itself, however, would not be consistent with the results of demographic analysis for these three Censuses (Statistics Canada 1987).

Remaining consistent with the demographic estimates, the differences would be explained in part by the presence of a large downward bias in the 1971 RRC estimate of persons missed. The 1971 unbiased estimate would have to be of the order of 3.8% rather than the estimated 1.9%. This would have to be accompanied by a not as large decrease in overcoverage between 1966 and 1971 followed by an increase in overcoverage for 1976 and a decrease for 1981. There would have to be also some under-estimation of missed persons for 1976.

Such a scenario is speculative, however, and no reason was found for such changes occurring. Other scenarios may also be possible. The occurrence of the discrepancies, however, does raise questions about the reliability of the RRC estimates and the potential effect of overcoverage on net coverage error.

The provincial distribution of the discrepancy between the RRC estimate of enumerated persons and the Census count differ among Censuses, further confounding its effects and potential sources. These results for the 1976 Census are given in Table 5.

Table 5
 Difference Between Reverse Record Check Estimates of Persons
 Enumerated and the 1976 Census Counts

Province	Difference in Population Enumerated (RRC-1976 Census)	Percent Difference
Canada (10 provinces)	-323,500	-1.4
Newfoundland	21,900	3.9
Prince Edward Island	-500	-0.4
Nova Scotia	- 4,500	-0.5
New Brunswick	-15,000	-2.3
Québec	-56,200	-0.9
Ontario	-207,000	-2.5
Manitoba	- 6,600	-0.6
Saskatchewan	1,400	0.1
Alberta	-43,400	-2.4
British Columbia	-12,800	-0.5

3.2 Variation in Geographic Distributions

The RRC estimates of undercoverage can be used as general indicators of the coverage quality of the Census. They are also intended to be used to direct the development and testing of coverage improvement procedures for future Censuses. Under ideal circumstances, they would be used to model undercoverage to produce estimates for small areas and as part of a coverage adjustment "correction" procedure. For these uses, geographic variation in coverage quality, indicated by the RRC results, is of particular concern. Variation in Census data distributions have been examined to determine whether they are correlated to the apparent variation in undercoverage among provinces. To date these investigations have not yielded satisfactory models or explanations.

A lack of success modelling undercoverage or explaining the variation between provinces may be due to, or confounded by: (i) bias and/or sampling error in the RRC estimates; (ii) undercoverage not strongly correlated to the Census characteristics of individuals, households and/or families; (iii) undercoverage correlated to a perhaps complex combination of Census and other characteristics; and/or (iv) a multitude of sources of undercoverage that must be considered separately; for example, undercoverage of individuals considered separately from undercoverage of entire households.

4. CONCLUSION

The RRC is thought to be the best vehicle developed to date for estimation of undercoverage in the Census in Canada. Its estimates provide basic measures to monitor and assess the quality of Census counts.

There are conceptual, theoretical and practical limitations to the RRC Check method as currently applied to the Canadian Census. The frames or lists used, while covering the large

majority of the population to be enumerated, are not comprehensive. Specific geographic areas are excluded as are certain segments of the population. The sample size is limited, but not necessarily to its present size, by constraints of tracing and matching, and by the demands for accuracy in operations. The "not traced" cases are a source of bias. The proportion of cases not traced, relative to the proportion of "missed" cases, in particular, adds an important uncertainty to the estimates, as does the inconsistency of RRC estimates of enumerated persons with corresponding Census counts.

In some instances the degree or impact of error, or limitations, could be evaluated in greater depth. Modifications and alternative procedures or methods that have a reasonable likelihood of improving the quality and applicability of the estimates can be applied. Potentially, alternatives can be developed. Such changes, however, would have varying costs and degrees of effectiveness associated with them. Also, it remains to be shown whether such changes would do more than enhance the status of the RRC estimates as general indicators of coverage quality in the Census.

ACKNOWLEDGEMENTS

The author would like to thank Gordon Brackstone, Geoff Hole, Judy Clarke and staff of Social Survey Methods for assistance and comments during preparation of the paper. The comments of the referees and editors were helpful and appreciated.

Appendix 1

Further Results From the Reverse Record Check

Results from the 1986 Census Reverse Record Check have been published (Statistics Canada 1988). The following extract displays the undercoverage rates for the 1981 and 1986 Censuses for demographic characteristics. Analysis of the 1986 undercoverage estimates by province, age, sex, marital status, mother tongue and other groupings is continuing.

1981 and 1986 Reverse Record Check Undercoverage Rates for Selected
Population Characteristics - 10 Provinces

Characteristic	1981 Estimated Population Undercoverage		1986 Estimated Population Undercoverage	
	Rate	S.E.	Rate	S.E.
	%	%	%	%
Sex				
Male	2.37	0.13	3.91	0.16
Female	1.65	0.12	2.87	0.16
Age Group				
0- 4	1.21	0.22	2.28	0.48
5-14	1.23	0.21	2.12	0.26
15-19	2.96	0.52	3.89	0.60
20-24	5.51	0.29	9.06	0.45
25-34	2.31	0.28	4.76	0.32
35-44	2.20	0.26	2.40	0.32
45-54	0.81	0.23	1.77	0.28
55-64	0.91	0.29	2.09	0.31
Marital Status				
Married/Separated	1.22	0.11	1.89	0.15
Divorced	5.10	1.03	7.07	1.07
Widowed	0.64	0.39	2.68	0.51
Single/Never Married	2.86	0.16	4.91	0.21
Mother Tongue				
English	1.86	0.11	3.12	0.13
French	1.80	0.20	3.10	0.33
Other	3.08	0.26	-	-
Urban/Rural Population Size Group				
Urban Areas	2.08	0.11	3.28	0.13
500,000 & over	2.29	0.17	3.58	0.15
100,000 to 499,999	1.86	0.31	2.94	0.33
Less than 100,000	1.80	0.23	-	-
Rural Areas	1.79	0.21	3.73	0.29

Appendix 2

Equations Used to Assess RRC Estimates and Estimator

The 1981 Reverse Record Check estimates have been assessed and discussed based upon four equations. The first simply defines the RRC population or frames. The second redefines the RRC sample in terms of the outcome or estimates of the study. The third defines the population enumerated in the Census in terms of the RRC estimate of enumerated persons. The fourth defines the error components for the estimate of missed persons.

Equation 1:

The RRC population size = $C_{76} + M_{76} - e(\hat{M}_{76}) + I_{76/81} + B_{76/81}$,

where

- C_{76} = number of persons counted, or enumerated, in one of the ten provinces in the 1976 Census,
- M_{76} = number of persons missed in one of the ten provinces in the 1976 Census,
- $e(\hat{M}_{76})$ = error (under or (-) over estimation of persons) associated with M_{76} , the Missed frame sample; *i.e.*, $M_{76} = \hat{M}_{76} + e(\hat{M}_{76})$,
- $I_{76/81}$ = number of registered 1976 to 1981 intercensal immigrants to one of the ten provinces,
- $B_{76/81}$ = number of registered 1976 to 1981 intercensal births in one of the ten provinces.

Equation 2:

The RRC estimates = $\hat{C}_{81} + \hat{C}_{fT81} + \hat{M}_{81} + \hat{M}_{fT81} + \hat{L}_{76/81} + \hat{A}_{81} + \hat{D}_{76/81} + \hat{O}_{f81}$

where

- \hat{C}_{81} = estimated number of persons in an RRC frame who were enumerated in one of the ten provinces in the 1981 Census,
- \hat{C}_{fT81} = estimated number of persons in an RRC frame who were enumerated in one of two territories in the 1981 Census,
- \hat{M}_{81} = estimated number of persons in an RRC frame who were missed in one of the ten provinces in the 1981 Census,
- \hat{M}_{fT81} = estimated number of persons in an RRC frame who were missed in one of the two territories in the 1981 Census,
- $\hat{L}_{76/81}$ = estimated number of persons in an RRC frame who were 1976 to 1981 intercensal emigrants,
- \hat{A}_{81} = estimated number of persons in an RRC frame who were abroad and had no usual place of residence in Canada at the time of the 1981 Census,
- $\hat{D}_{76/81}$ = estimated number of persons in an RRC frame who died in the 1976 to 1981 intercensal period,
- \hat{O}_{f81} = estimated overcoverage (number of "persons") in the Census, Birth and Immigrant frames which was detectable in the 1981 RRC operations.

Equation 3:

The estimate \hat{C}_{81} should = $C_{81} - \hat{C}_{81}[e(\hat{M}_{76})] - \hat{C}_{c/re81} - R_{81} - T_{76/81} - S_{76/81} + M_{n81} - O_{81} + \hat{C}(\hat{O}_{nf81})$,

where

- C_{81} = number of persons enumerated in one of the ten provinces in the 1981 Census,
 $\hat{C}_{81}[e(\hat{M}_{76})]$ = that component of $e(\hat{M}_{76})$ not or (-) over represented in \hat{C}_{81} ,
 $\hat{C}_{c/re81}$ = under or (-) over-estimation of "enumerated" persons in an RRC frame because of classification, response, sampling and "no trace" error in the 1981 RRC,
 R_{81} = number of persons abroad at the time of the 1976 Census who were in Canada at the time of the 1981 Census,
 $T_{76/81}$ = number of intercensal migrants from the two territories to a province,
 $S_{76/81}$ = net number of intercensal entries to the ten provinces, as of Census Day, not in an RRC frame and not accounted for above (e.g., illegal aliens),
 M_{n81} = number of persons not in a RRC frame who were missed in one of the ten provinces in the 1981 Census,
 O_{81} = overcoverage in the ten provinces in the 1981 Census,
 $\hat{C}(\hat{O}_{nf81})$ = estimated overcoverage (number of "persons") in the Census, Birth and Immigrant frames which was *not detected* in the 1981 RRC operations and is represented in \hat{C}_{81} .

Thus,

$$\hat{C}_{81} - C_{81} = -\hat{C}_{81}[e(\hat{M}_{76})] - \hat{C}_{c/re81} - S_{76/81} + M_{n81} - O_{81} + \hat{C}(\hat{O}_{nf81}) - R_{76/81} - T_{76/81},$$

assuming no error in \hat{O}_{nf81} .

Equation 4:

$$M_{81} - \hat{M}_{81} = \hat{M}_{81}[e(\hat{M}_{76})] - \hat{M}_{81}(\hat{O}_{nf81}) + \hat{M}_{c/re81} + M_{n81} = e(\hat{M}_{81}).$$

where

- $\hat{M}_{c/re81}$ = under or (-) over-estimation of "missed" persons in an RRC frame because of classification, response, sampling and "no trace" error in the 1981 RRC,
 $\hat{M}_{81}[e(\hat{M}_{76})]$ = that component of $e(\hat{M}_{76})$, represented in \hat{M}_{81} ,
 $\hat{M}_{81}(\hat{O}_{nf81})$ = estimated overcoverage (number of "persons") in the Census, Birth and Immigrant frames which was not detected in the 1981 RRC operations and is represented in \hat{M}_{81} .

Note: There is a classification, response, sampling and "no trace" error component associated with each item of equation 2; e.g., $\hat{C}_{c/re81}$ and $\hat{M}_{c/re81}$. These taken in total sum to zero. In the above equations these error components exclude error caused by overcoverage and overcoverage which results in a "not traced"; e.g., non-existent persons enumerated in the previous Census. The effect of overcoverage is included, for example, in $\hat{C}(\hat{O}_{nf81})$ and $\hat{M}(\hat{O}_{nf81})$.

Similarly,

$$e(\hat{M}_{76}) = \hat{M}_{76}[e(\hat{M}_{71})] - \hat{M}_{76}(\hat{O}_{nf76}) + \hat{M}_{c/re76} + M_{n76}.$$

Error and part of the difference $\hat{C} - C$ can be passed from one RRC to another through the Missed frame and through overcoverage in the Census frame. This error could account for a large part of the difference $\hat{C}_{81} - C_{81}$. The effect on $\hat{C} - C$ may be much greater than on \hat{M} .

The rate of net coverage error in the 1981 Census, for the ten provinces, would be equal to:

$$\frac{M_{81} + M_{n81} - O_{81}}{C_{81} + M_{81} + M_{n81} - O_{81}};$$

and the rate of undercoverage would be:

$$\frac{M_{81} + M_{n81}}{C_{81} + M_{81} + M_{n81} - O_{81}}.$$

The estimator used in the RRC is

$$\frac{\hat{M}_{81}}{C_{81} + \hat{M}_{81}}.$$

Even a relatively small value of $M_{n81} - O_{81}$ could contribute significant bias to the results of the RRC, if these results are used as estimates of net coverage error. A relatively small value of $e(\hat{M}_{81})$ could contribute significant bias to the RRC undercoverage estimates: two potential elements of bias coming from the previous RRC; one from any misclassification within the RRC; and one from “missed” persons among those not included in an RRC frame. There may be, of course, some cancellation among these elements.

An alternative estimator would be to use \hat{C}_{81} instead of C_{81} in the denominator. There are specific and not unlikely circumstances under which the use of C_{81} would produce estimates with less bias at the national level. These circumstances, which involve the relative sizes of $\hat{C}_{81} - C_{81}$, O_{81} and M_{n81} do not hold, however, for provinces or estimates for which the Census count of enumerated is less than the RRC estimate of enumerated.

REFERENCES

BURGESS, R.D. (1986). Major issues and implications of tracing survey respondents. International Symposium on Panel Surveys, Washington D.C.

GOSSELIN, J.-F. (1976). The Methodology of the 1971 Reverse Record Check. *Survey Methodology*, 2, 180-193.

PLAUT, G. (1985). Refugee determination in Canada. House of Commons Standing Committee on Labour, Manpower and Immigration.

STATISTICS CANADA (1976 to 1981). *Vital Statistics Volume III: Death/Mortality Catalogue* 84-206, Statistics Canada.

STATISTICS CANADA (1979). *Mobility Status and General Population Characteristics*. Catalogue 92-834, Statistics Canada.

- STATISTICS CANADA (1982). *Population Counts - 1976 and 1981 - Federal Electoral Districts*. Catalogue 99-908, Statistics Canada.
- STATISTICS CANADA (1983a). *Mobility Status*. Catalogue 92-907, Statistics Canada.
- STATISTICS CANADA (1983b). Reverse Record Check Tabulations, 1981 Data Quality Project, 1981 Census of Population, unpublished.
- STATISTICS CANADA (1984). Public Use Reverse Record Check Tape Users Guide. 1981 Census of Population, Statistics Canada.
- STATISTICS CANADA (1987). *Population Estimation Methods, Canada*. Catalogue 91-528E, Statistics Canada.
- STATISTICS CANADA (1988). Undercoverage Rates from the 1986 Reverse Record Check. User Information Bulletin, Number 2, Statistics Canada.

A Demographic Approach to the Evaluation of the 1986 Census and the Estimates of Canada's Population

ANATOLE ROMANIUC¹

ABSTRACT

A significant increase in coverage error in the 1986 Census is revealed by both the Reverse Record Check and the demographic method presented in this paper. Considerable attention is paid to an evaluation of the various components of population growth, especially interprovincial migration. The paper concludes with an overview of two alternative methods for generating postcensal estimates: the currently-in-use, census-based model, and a flexible model using all relevant data in combination with the census.

KEY WORDS: Census undercoverage; Population estimates; Demographic component method.

1. INTRODUCTION

The accuracy of the census, and of the postcensal population estimates based thereon, is an important issue in its own right. The use of population numbers in the formulae for calculating revenue transfers between various levels of government, makes the question of accuracy all the more critical and politically sensitive (Fellegi 1980; Romaniuc and Raby 1980). The intense debates on whether or not to adjust population counts for census undercoverage in Canada and the USA, and several judicial litigations fought in the latter country in recent years, are indications of both the political importance and the technical complexity of the issue.

Yet, in spite of all that has been written on the subject, the elaborate arguments marshalled by both those for and those against adjustment, the debates remain inconclusive (Keyfitz 1979 and 1981; Kish 1980; Spencer 1980; Freedman and Navidi 1986; Stoto 1987). Eventually Statistics Canada decided (as did the US Department of Commerce) against adjustment for census undercoverage, while at the same time reaffirming its long-standing commitment to the policy of data quality evaluation (Wilk 1981). By making public both the evaluation results and the underlying methodology, the users can make adjustments to suit their particular needs, in full knowledge of the strengths and limitations of the census counts and estimates. It is in the spirit of this policy on quality evaluation that this paper has been written.

There are basically two approaches to the evaluation of the accuracy of census counts. One is the "micro" approach, involving individual verification, case-by-case record matching, in order to identify persons who have been missed, enumerated more than once, or enumerated even though, by definition, they are not part of the census universe. To this type of evaluation belong the US Bureau of the Census Post-Enumeration Program and Statistics Canada's Reverse Record Check (RRC).

¹ Anatole Romaniuc, Director, Demography Division, Statistics Canada, Ottawa, Ontario.

The second is the “macro” evaluation approach involving an analysis at aggregate levels, such as comparison of the census counts with figures derived from independent sources or with estimates arrived at by means of statistical and demographic methods. Following the pioneering work by Ansley Coale (1955), the demographic techniques of analysis have been used by the US Bureau of the Census to evaluate census coverage concurrently with the Post-Enumeration Program (see most recent report by Fay, *et al.* 1988). Some earlier attempts of this kind in Canada were also made (Lapierre 1970). The essence of the demographic method, as we shall see later, is that it brings to bear the formal relationship between population and its growth components – namely births, deaths and migration.

The evaluation of the 1986 Census coverage through the Reverse Record Check (RRC) has been carried out and reported upon elsewhere (Carter 1988; Statistics Canada 1988). It suffices to say that the RRC-based estimates of undercoverage are subject to sampling error – which can be quite significant for provinces with a small population – and to biases of unknown magnitudes (difficulties in tracing persons or matching individual records). Furthermore, the RRC has been designed primarily to measure undercoverage. The measurement of overcoverage has been attempted on an experimental basis, but at the time of writing, the results were unavailable. For these and similar reasons, an alternative assessment of the accuracy of the census counts becomes all the more important.

This paper evaluates, by means of demographic analysis, the accuracy of the three most recent censuses, with emphasis on the 1986 Census. A three-step operation is followed. First, census counts and population estimates are compared with each other. Second, demographic techniques are used to generate alternative estimates of census undercoverage which are, in turn, compared with those based on the Reverse Record Check. As a third and final step, the focus of evaluation is shifted from census counts to intercensal change in population. Two sets of independent estimates of intercensal population change are produced. One is based on the two consecutive censuses, while the other is obtained directly from data on births, deaths and migration.

Before proceeding with the actual evaluation, a word of caution is in order. Though of acceptable quality for most of the uses they serve, neither census counts nor population estimates are perfect. Indeed, there is no one set of data deemed to be perfect enough to serve as a benchmark for the validation of other data. The statistical reality is that data are imperfect in varying degrees. The fine tuning and high precision that would be required for particular uses – such as government allocations and revenue transfers referred to earlier – might not be attainable under the present state of the art. However, we hope that this evaluation, using a combination of statistical tools, imperfect as they may be, will enable us to get some sense of the direction and magnitude of errors and biases affecting census population counts and various components of population estimates. Such an undertaking will hopefully set the stage for improvements as we work toward the 1991 Census and the post-1991 population estimation methodology.

2. CENSUS COUNTS VERSUS POPULATION ESTIMATES: ERROR OF CLOSURE

The postcensal estimates of population are obtained, as per equation 1, by the so-called component method, whereby births and immigrants are added to, and deaths and emigrants are subtracted from, the base census population. The net interprovincial migration is then added to estimate population by province. The procedure is repeated annually over the five-year period

to the next census. The current estimation methodology calls for postcensal estimates to be revised retrospectively so as to bring them in line with the latest census counts (Statistics Canada 1987). The difference, as per equation 2, between estimates thus arrived at and census counts is termed "the error of closure" (EC).

$$\hat{P}_t = R_{t-5} + \left[B_{t-5,t} - D_{t-5,t} + I_{t-5,t} - \hat{E}_{t-5,t} + \hat{N}_{t-5,t} \right] \quad (1)$$

$$EC (\%) = \frac{\hat{P}_t - R_t}{R_t} \times 100, \quad (2)$$

where:

- \hat{P}_t = estimated population at time t ;
- R = census counts at time t or $t-5$ as the case may be;
- B = number of births;
- D = number of deaths;
- I = number of immigrants;
- \hat{E} = number of emigrants as estimated;
- \hat{N} = net interprovincial migration as estimated;
- $t-5, t$ indicates the five-year period during which the events occurred.

Table 1 presents the error of closure for the last four censuses for Canada, provinces and territories. On the whole, agreement between the census counts and the population estimates is fairly good even for provinces. This is all the more remarkable considering the fact that, in the absence of direct records, both emigration from Canada and interprovincial migration have to be estimated from administrative data (family allowance and income tax files).

Despite the high level of agreement, there are two salient features in the error of closure. One such feature is the jump to nearly one percent error of closure in 1986, a relatively large error when compared to that in the previous censuses. For the 1971 and 1976 censuses the error stood at slightly over one-half of one percent and only at one-quarter of one percent in 1981. The other feature is the negative error of closure in 1981. Whereas in the other three censuses, the estimates exceeded the census counts, in 1981 the former fell short of the latter. Almost all of this shortfall originated in the province of Alberta.

Turning to the provinces, one notes a consistently positive error of closure in 1986, whereas the sign of the error varied in the previous three censuses. Furthermore, for most of the provinces, the magnitude of the error has increased in 1986 as compared to the previous three censuses. The larger errors of closure were found in the Maritime Provinces and Quebec, and the smaller in Ontario and in the Western Provinces, with the exception of Saskatchewan.

The 1981 case of Alberta, referred to above, calls for some further remarks. In 1981, this province had to contend with an unusually large negative error of closure: the estimates fell short of the census count by 53,886 individuals or 2.41%. There are two possible explanations for this outcome. One is that the 1981 Census in this province may have suffered from a

Table 1
Error of Closure: Canada, Provinces and Territories,
June 1971, 1976, 1981 and 1986

Geographic Area	Percent Error ¹			
	1971	1976	1981	1986
Canada	0.51	0.58	-0.25	0.95
Newfoundland	0.32	-0.19	1.25	2.02
Prince Edward Island	-0.76	1.58	-0.31	1.06
Nova Scotia	-2.45	0.93	-0.03	1.28
New Brunswick	-0.44	1.51	-0.28	1.57
Quebec	0.08	0.10	-0.58	1.34
Ontario	1.41	1.07	0.37	0.73
Manitoba	-0.01	1.21	0.83	0.57
Saskatchewan	0.21	0.91	-0.52	1.06
Alberta	0.31	-0.09	-2.41	0.81
British Columbia	0.47	0.07	-0.22	0.58
Yukon	-6.63	-2.34	-2.11	-4.66
Northwest Territories	3.14	-0.92	-5.60	-1.32

¹ $\frac{\text{Population Estimate} - \text{Census Count}}{\text{Census Count}} \times 100$

Source: Demography Division, Statistics Canada.

relatively large “overcount”. Prompted by the booming oil-based economy, a great number of transient job-seekers from other provinces made their way to Alberta, some of whom may have been incorrectly enumerated as this province’s usual residents. Yet, the fact that for 1981 Alberta showed an above-average undercount (2.54%) only adds to the puzzle. The other possible explanation is that the flow of in-migrants to Alberta, in those days of its economic prosperity and demographic boom, was not fully captured by the family allowance and taxation files – the basis of interprovincial migration estimates. In other words the large shortfall in the 1981 estimates of population might have resulted from an understatement of the net migration to Alberta.

Having demonstrated that the gap between estimates and counts widened significantly in 1986, the question to be addressed in the subsequent sections is whether this is due to the deterioration of: (a) the census coverage or (b) the data on the components of population growth over the last intercensal period.

3. DEMOGRAPHICALLY-DERIVED UNDERCOVERAGE RATE

By adjusting the census base population for undercoverage as estimated from the RRC, and by adding the net population increase (births, deaths and migrants) over the subsequent postcensal period, one obtains, as per equation 3, the population at the time of the next census. We shall call this the *expected* population, to differentiate it from the *estimated* and *enumerated* populations dealt with in the previous section.

$$P'_t = \left[R_{t-5} + \hat{U}_{t-5} \right] + \hat{G}_{t-5,t}, \tag{3}$$

where:

- P'_t = expected population at time t ;
- R_{t-5} = enumerated population at time $t-5$;
- \hat{U}_{t-5} = the number of individuals missed in the census $t-5$, as estimated through the Reverse Record Check (RRC);
- \hat{G}_{t-5} = estimates of net population change over the intercensal period $t-5, t$ (births, deaths and migrants in equation (1)).

The difference, U'_t , between the *expected population*, P'_t , and the *enumerated population*, R_t , as per equation 4, can be taken here as a coverage error. We shall call this the *demographic estimate of coverage error*.

$$U'_t = P'_t - R_t. \tag{4}$$

And the rate of coverage error, u'_t , is simply the ratio of the demographically estimated error of coverage, U'_t , to the expected population, P'_t :

$$u'_t = \frac{P'_t - R_t}{P'_t} = \frac{U'_t}{P'_t}. \tag{5}$$

For comparison, the undercoverage rate as estimated through the RRC stands as follows:

$$\hat{u}_t = \frac{\hat{U}_t}{R_t + \hat{U}_t}. \tag{6}$$

How do the demographically estimated error of coverage and the RRC-estimated undercoverage compare? First, it should be stressed that both are subject to error and bias. The former is affected by: (a) the lack of an estimate of overcoverage; (b) the biases in the RRC-based undercoverage \hat{U} at t and $t-5$ censuses, and; (c) the biases involved in the estimates of intercensal net population change $\hat{G}_{t-5,t}$, particularly its migration component. The RRC estimate of undercoverage is affected by: (a) sampling error, and; (b) various biases due to tracing of individuals, record matching, *etc.* Furthermore the undercoverage rate, \hat{u} , as per formula (6), is slightly downwardly biased because R_t in the denominator includes an overcount of unknown quantity. Hence, alone on these grounds, comparison between the two coverage measurements is far from being straightforward.

But there are conceptual differences as well. The RRC estimate is a pure undercoverage measurement. Demographically estimated coverage error is a more complex, difficult to define unequivocally, entity. It is neither an undercoverage nor a net undercoverage. In order, to better grasp the relationship between the two, the equation (3) of the expected population, P'_t , may be rewritten as per (7). Note that the enumerated population, R , is now expressed in terms of its two components: those who were correctly enumerated, R' , and those who were overcounted, O .

$$P'_t = \left[(R'_{t-5} + O_{t-5}) + \hat{U}_{t-5} \right] + \hat{G}_{t-5,t}. \tag{7}$$

The undercoverage rate estimated by the demographic method as expressed in equation (5) now becomes:

$$u'_t = \frac{[(R'_{t-5} + O_{t-5}) + \hat{U}_{t-5} + \hat{G}_{t-5,t}] - (R'_t + O_t)}{(R'_{t-5} + O_{t-5}) + \hat{U}_{t-5} + \hat{G}_{t-5,t}} \quad (8)$$

It follows from (8) that the overcoverage affects both the expected and the enumerated populations. Consequently, the demographic rate of undercoverage reflects the combined effect of the undercoverage per se and the difference in the overcoverage, 0, of the base census, $t-5$, and terminal census, t . Assuming that both (a) the RRC-based undercoverage, \hat{U} at t and $t-5$, and (b) the population change (the net sum of the components) for intercensal period, $\hat{G}_{t-5,t}$, are correctly estimated, then the demographic coverage rate, u'_t , and the RRC rate, \hat{u}_t , will vary numerically depending on the level of the overcoverage of censuses at time, $t-5$ and t , so that if $O_t \geq O_{t-5}$ then $\hat{u}_t \leq u'_t$.

Having clarified the conceptual particularities of the two measures of coverage error, we now turn to Table 2 which presents for Canada the coverage estimates for the 1981 and 1986 censuses. Both estimates reveal a significant increase in the coverage error in the 1986 Census. However, the demographically-derived rate of coverage error is consistently lower than the RRC rate of undercoverage: 2.82% and to 3.21% for 1986, and 1.70% and 2.01%, for 1981, respectively. This could mean that the overcoverage was higher in 1981 than in 1976, and higher in 1986 than in 1981, on the condition that the assumptions underlying the identities are correct. But there are no data to either confirm or deny the validity of these assumptions.

The estimates of coverage error by the two methods – demographic and RRC – by province in Table 2 are portrayed by Figure 1(a) and 1(b). The explanation of the differences at the provincial level is liable to present even greater uncertainties because the error and biases,

Table 2
Demographic and Reverse Record Check Estimates of Undercoverage Rates:
By Provinces, 1981 and 1986

Geographic Area	Demographic Method		Reverse Record Check ¹			
	1981 (%)	1986 (%)	1981 (%)		1986 (%)	
Canada (Territories not included)	1.70	2.82	2.01	(0.09)	3.21	(0.12)
Newfoundland	2.29	3.60	1.74	(0.95)	2.01	(0.32)
Prince Edward Island	0.05	2.10	1.17	(0.54)	2.16	(0.80)
Nova Scotia	0.82	2.22	1.05	(0.34)	2.63	(0.38)
New Brunswick	1.83	3.28	1.81	(0.30)	2.83	(0.36)
Quebec	2.31	3.13	1.91	(0.21)	3.06	(0.29)
Ontario	1.81	2.53	1.94	(0.14)	3.40	(0.19)
Manitoba	1.88	1.44	0.98	(0.35)	2.22	(0.40)
Saskatchewan	0.76	2.00	0.99	(0.37)	2.51	(0.36)
Alberta	-1.18	3.09	2.54	(0.36)	2.75	(0.33)
British Columbia	2.62	3.55	3.16	(0.33)	4.49	(0.39)

¹ Figures in brackets are Standard Deviations.

Source: Demography Division, Statistics Canada.

referred to above, at these levels are expected to be larger than they are at the national level. This is true in particular for sampling error in the case of the RRC undercoverage estimates, and for the biases in the interprovincial migration affecting net intercensal population change in the case of the demographic estimates of coverage error.

With the above comments regarding the biases and conceptual differences in mind, let us see how consistent are the two coverage measures at the provincial level? To this end, the following criterion of consistency is posited: if the two measures of coverage were conceptually identical and empirically correct, their respective correlation points in space should line up along the 45° bisectrix.

For the 1981 Census, disregarding the special case of Alberta referred to earlier (and also P.E.I. heavily affected by the sampling error), the correlation points follow closely the theoretical 45° straight line. The discrepancies are small: in most cases they are not statistically significant given the standard deviation affecting the RRC estimates (see Table 2).

For the 1986 Census, six provinces out of ten (Saskatchewan, Nova Scotia, Prince Edward Island, Quebec, Alberta and New Brunswick) have their respective points falling within close range of the 45° bisectrix and thus meet the consistency test. One, Newfoundland, falls far afield on the left side, suggesting a possible understatement of the RRC undercoverage rate for this province. Manitoba, Ontario and British Columbia fall well to the right side of the 45° bisectrix suggesting a possible overstatement of the RRC undercoverage or understatement of demographic coverage rate.

It should be stressed once again that the analysis of the accuracy of census coverage has been hampered by the lack of information on overcoverage. Yet, it is fair to say that notwithstanding its limitations, the analysis strongly points to a deterioration of the 1986 census coverage.

4. CENSUS AND COMPONENT-BASED INTERCENSAL POPULATION CHANGE: A CHECK FOR CONSISTENCY

The task now at hand is to compare two sets of *independent* estimates of the intercensal net population change: one set based on demographic components (births, deaths and migration), the other set derived from two consecutive censuses, unadjusted and adjusted for undercoverage. Refer to the former as *component-based estimates* and to the latter as *census-based estimates* of intercensal net population change.

$$\hat{G}_{t-5,t} = B_{t-5,t} - D_{t-5,t} + I_{t-5,t} - \hat{E}_{t-5,t} + \hat{N}_{t-5,t} \quad (9)$$

$$\bar{G}_{t-5,t} = R_t - R_{t-5} \quad (10)$$

$$\bar{G}'_{t-5,t} = (R_t + \hat{U}_t) - (R_{t-5} + \hat{U}_{t-5}). \quad (11)$$

All the above notations have been made explicit in the previous formulae.

Two independently-produced estimates might be construed as reasonably trustworthy if they are similar for a given point in time. As seen in Table 3, the difference between census-based and component-based estimates is only about 5% for the 1976-81 period. For the 1981-86 period, the two estimates differ by a substantial margin of 19% if unadjusted, and by 8% if adjusted for undercoverage.

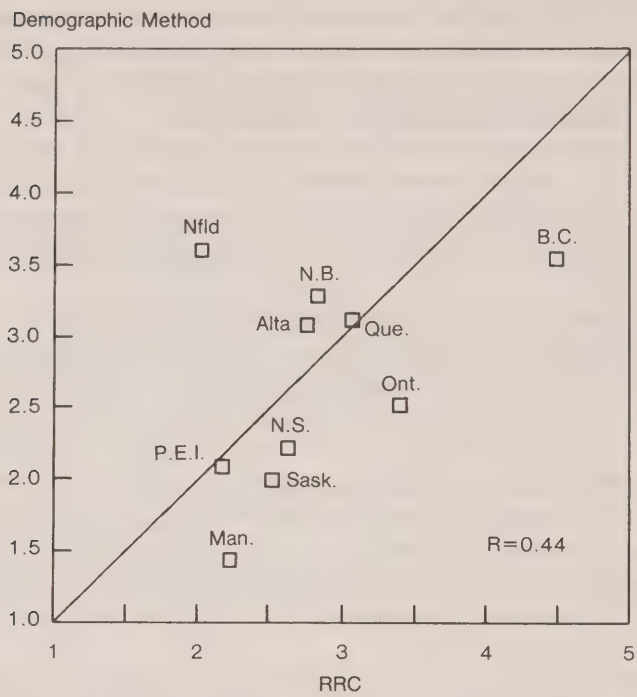


Figure 1. Relationship between Undercoverage Rates as Estimated by Reverse Record Check and Demographic Method, 1986 Census

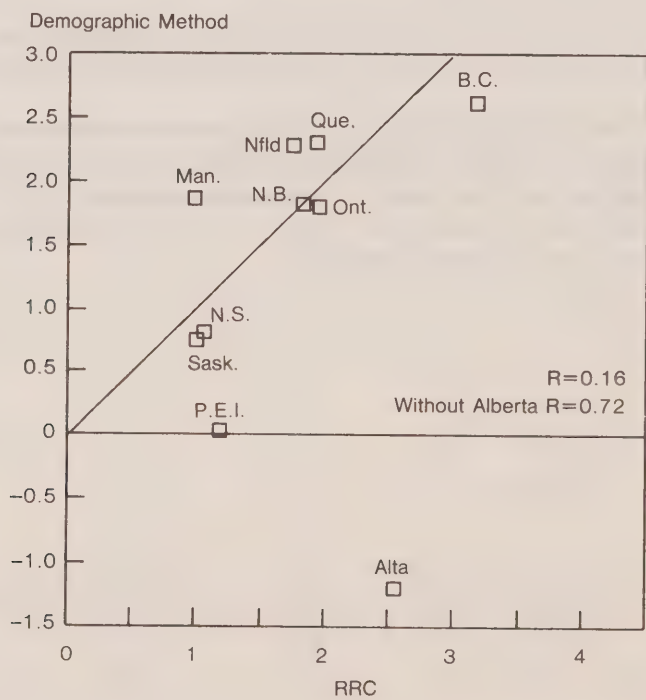


Figure 2. Relationship between Undercoverage Rates as Estimated by Reverse Record Check and Demographic Method, 1981 Census

Table 3
Ratio Between Census and Component-Based Intercensal Change in Population:
By Province, 1976-81 and 1981-86

Geographic Area	Ratio between Census and Component-based Intercensal Population Change Multiplied by 100			
	1981-86		1976-81	
	Not adjusted for Census Undercoverage	Adjusted for Census Undercoverage	Not adjusted for Census Undercoverage	Adjusted for Census Undercoverage
Canada	80.9	108.4	104.5	106.1
(Territories not included)				
Newfoundland	5.7	19.6	58.3	80.9
Prince Edward Island	76.7	101.6	109.8	135.7
Nova Scotia	70.4	110.3	101.3	111.1
New Brunswick	55.6	86.7	111.3	99.2
Quebec	54.2	97.1	122.7	83.8
Ontario	88.1	115.2	92.0	103.2
Manitoba	89.2	117.3	35.6	29.0
Saskatchewan	79.4	110.3	112.0	105.5
Alberta	88.8	94.5	115.6	124.4
British Columbia	89.5	118.3	102.2	105.8

Note: The procedure cannot be applied for the period 1971-76 because, for this and earlier periods, emigration has been estimated residually from the two consecutive censuses and the remaining growth components (births, deaths and immigrants).

Source: Demography Division, Statistics Canada.

The comparison by province is a more delicate matter. On the components side, one has to contend with the reliability of the interprovincial migration estimates. On the census side, one must reckon with the variability of biases in undercoverage and overcoverage, and sampling errors in the RRC undercoverage estimates. Sampling errors alone could account for up to 15% of variations in the ratio between the two estimates of the intercensal population change for some provinces. Any variations beyond this level are more likely to have been induced by errors and biases from other than the sampling.

Hence, in the absence of a more trustworthy criterion, we have set $\pm 15\%$ as a tolerance limit for the discrepancies between the two estimates. The tolerance limit thus set, has at least the merit of screening out highly questionable cases.

With these qualifications in mind, let's turn to Table 3, which compares by province, census and component-based population changes for the last two intercensal periods. Six provinces out of ten for the 1976-81 period, and four out of ten for the 1981-86 period meet the somewhat arbitrarily set tolerance test. In general, the discrepancies are wider for the 1981-86 period than for 1976-81. Particularly conspicuous in this regard are the provinces of Newfoundland, Quebec, and New Brunswick.

Newfoundland's census-based 1981-86 population change represents only 5% of that derived from the components. It is still only 19%, even after adjustment for undercoverage. Such a low population growth would call for a net migration loss of about 26,000 over the 5-year period. Yet, all the three sources of interprovincial migration (Family Allowance, Taxation and the census mobility question) place these losses in the range of 14,800 to 16,500 (see Table 5).

Similar inconsistencies are found in the case of Quebec. The census-based population growth for the period 1981-86, which represents only 64% of the component-based growth, would imply Quebec's loss through out-migration to be twice the amount estimated by Statistics Canada, that is, 160,000 instead of 80,000. Yet again, all the three sources of information put the net-migration losses in the range of 63,000 to 81,000 over the 5-year period. The gap between the two estimates of intercensal change is almost wiped out when the 1981 and 1986 census counts are adjusted for undercoverage.

The case of New Brunswick is similar to that in Quebec and Newfoundland. The census-based estimate of population growth for the 1981-86 period suggests a net loss through out-migration of 11,200, whereas the family allowance-based figure is 2,200. The census mobility question and taxation figures are even lower, 1,376 and 65, respectively. Adjustment for undercoverage would bring New Brunswick's two estimates of the intercensal population change well within the tolerance limit.

What, then, can be concluded from the above analysis regarding the intercensal population change? It appears that both the components and the census generate reasonably consistent estimates of population change for the 1976-81 period. The discrepancies are small, within a tolerable limit for Canada and for most of the provinces. This, however, is not the case for the most recent intercensal period, 1981-86. Something seems to have deteriorated and the question remains as to whether it is the census or the components of population growth. As was seen in the preceding section, the 1986 Census experienced a significant increase in undercoverage estimated by two different methods. Adjustment for undercoverage, however, did not always produce better estimates of intercensal population growth, in fact the opposite happened in some cases. In the next section, we take a closer look at the components of population growth.

5. HOW GOOD ARE THE COMPONENTS OF POPULATION GROWTH?

What follows is a brief assessment of the quality of the data on births, deaths, immigration, emigration, and interprovincial migration. For a more complete account of the data on those components, and methodologies for estimating migration, the reader is referred to the 1987 Statistics Canada publication "Population Estimation Methods, Canada".

The registration of births and deaths is deemed to be complete in this country. Deaths or births that somehow escape registration must be by necessity very small in number in view of the prevailing regulations (need for a burial certificate) and the material (family allowance) incentives and legal requirements for registering births. Some late registration may occur, but the numbers are small. For the 1981-85 period, 3,831 or 0.02% of all births and 2,528, or 0.03%, of all deaths were registered beyond the cut-off date. This makes a net of only 1,303 persons unaccounted for in the population estimates.

Immigration statistics are regarded as reasonably accurate to the extent one speaks here of landed immigrants. The distribution of immigrants by province is based on their intended destination rather than on where they actually settle. It is, however, noteworthy, as per Table 4, that this distribution closely agrees with the 1986 Census distribution of immigrants.

Compared to the three other components reviewed above – births, deaths and immigration – interprovincial migration and emigration are weaker links in equation (1) which is used for estimating population for postcensal years. There are indeed no direct records of internal migration or emigration. Such figures must be estimated indirectly from administrative files

Table 4
Percentage Distribution of Immigrants by Province Based on the 1981 Census
and Immigration Records of Intended Destination in 1980

Geographic Area	Immigration Records	Census
Newfoundland	0.4	0.3
Prince Edward Island	0.1	0.1
Nova Scotia	1.1	1.0
New Brunswick	0.8	0.8
Quebec	15.7	15.0
Ontario	43.5	42.7
Manitoba	5.4	5.4
Saskatchewan	2.5	2.6
Alberta	13.2	14.5
British Columbia + Yukon		
+ Northwest Territories	17.2	17.6
Canada	100.0	100.0

Source: Demography Division, Statistics Canada.

- family allowance and income tax - which contain information on changes of residence. They deserve, therefore, more than a cursory consideration. In what follows, we shall focus on the significant methodological and data improvements achieved in recent years, as well as address certain persistent shortcomings inherent to these estimates. For a more complete account see Chapters IV and V of the Population Estimation Methods, Canada, 1987.

While family allowance data have been used since 1956, the most significant innovation to the system for estimating interprovincial migration was the addition of personal income tax data in 1976. As of 1981, a "two-track" estimation system was implemented: the *preliminary* quarterly and annual estimates based on family allowance data, and the *final* annual estimates based on taxation data. Both these data sources have strengths and weaknesses.

The main advantage of the family allowance file lies in its timeliness and fairly high accuracy. The information on change of address is available two months after the fact. The accuracy of the file is contingent upon two factors. The first is the comprehensiveness of coverage of child population, as every child under 18 years of age, supported by a parent, is entitled to a monthly payment. The second is the financial incentive for the beneficiaries of family allowances to report any change of address as soon as it occurs. The family allowance file does not, however, provide information on adult migration. This has to be estimated indirectly, by applying a conversion factor, "*f*", which is obtained by calculating the ratio of the adult migration rate to the child migration rate from the taxation data available for the most recent year.

Given the key importance of the *f* factor in the estimation formulae, a few comments are called for. Prior to 1971, the value of *f* was based on 5-year migration data from the most recent census. As the annual age-specific data on migrants became available from income tax records, the decision was made to use such data since they have an advantage over census data in that they reflect a more recent age pattern of migration.

Another innovation is worth mentioning. Prior to 1981, the *f* factor was calculated only by province of origin. However, with the availability of relevant data from taxation, it became evident that this factor also varies significantly by province of destination. Consequently, the decision was made to calculate the *f* factor by both province of origin and province of destination.

Turning now to the personal income tax file as the data source for estimating interprovincial migration, the following assessment is in order. As compared to the family allowance file, the taxation file has the advantage of having a much broader demographic base: tax filers and their dependents represent roughly 90% of the population. However, there are various sources of potential errors and biases. Information on tax filers' dependents must be imputed from the dollar value of total exemptions. Various assumptions have to be made in imputing the migratory status of the tax filers' dependents, as well as that of persons who are neither filing income tax returns, nor are dependents upon those who do so, and therefore are not covered at all by the taxation system. This is particularly the case for young adults and the elderly, who may be more prone to neglect to file their tax-return or who may not earn the minimum income required for filing. Such differential age-related biases, if indeed present, affect the estimates of the age structure, and this in turn affects the value of the f factor, used in the family allowance-based preliminary estimates of interprovincial migration.

Table 5 presents figures on net interprovincial migration for the intercensal 1981-86 period based on family allowance, taxation, and the census question on residence five years ago. Notwithstanding some significant variations in numbers, the three sources of data provide a consistent picture of level of interprovincial net migration over the 5-year period, by province.

What has been said about interprovincial migration also holds for emigration – Canadians taking residence in another country. Prior to 1981, the aggregate emigration to countries other than the United States and the U.K. (for which data were available through the immigration services of the two countries) had to be estimated residually from consecutive censuses and the components of intercensal population growth. As of 1981, the estimation of the number of emigrants has been based on family allowance and income tax data. The procedure is similar to that described above for estimating interprovincial migration. Child-migration is estimated from family allowance data. To estimate adult emigration, and hence total emigration, a conversion factor, f , based on income tax data, is applied to child-emigration. This same procedure applies to both the preliminary and final estimates of emigration, except that in the latter case more complete data are used.

Table 5
Net Interprovincial Migration for the Period 1981-1986,
Based on Specified Sources

Geographic Area	1986 Census ¹	Family Allowance	Income Tax
Canada	0	0	0
Newfoundland	-16,550	-14,837	-15,051
Prince Edward Island	1,540	293	751
Nova Scotia	6,275	5,204	6,895
New Brunswick	-1,370	-2,239	-65
Quebec	-63,295	-76,040	-81,254
Ontario	99,355	115,497	121,767
Manitoba	-1,555	-3,700	-2,634
Saskatchewan	-2,820	-668	-2,974
Alberta	-27,665	-34,073	-31,676
British Columbia	9,500	13,289	7,382
Yukon	-2,665	-2,381	-2,775
Northwest Territories	-755	-345	-366

¹ Population of 5 years and over.

Source: Demography Division, Statistics Canada.

Table 6
Estimates of Emigrants by Different Methods, Canada, 1981-86

Method	1981-86
Residual Method from Censuses	
(a) Unadjusted for Undercoverage	476,406
(b) Adjusted for undercoverage	134,807
Revenue Canada Tax File	165,272
Family Allowance Method (current) (using the <i>f</i> factor from the tax file)	235,481
Family Allowance Method (proposed) (using the <i>f</i> factor from the immigration file)	275,762
Reverse Record Check ¹	288,376

¹ Preliminary.

Source: Demography Division, Statistics Canada.

Table 6 compares, for the 1981-86 intercensal period, the estimates of emigration based on the family allowance files with the estimates produced by the various alternative methods. Note that the residually-derived emigration estimates, whether from adjusted or unadjusted census counts, are out of line with the more plausible estimates derived from the administrative files and the Reverse Record Check (RRC).

In brief, significant enhancements have been made to the system used to estimate interprovincial migration and emigration, particularly since 1981. While it can be surmised that the overall quality of the estimates has improved as a result, no demonstrable proof can be adduced. The family allowance and income tax data are fraught with various shortcomings inherent in any data system that has been designed for administrative rather than for statistical purposes.

6. CONCLUSIONS AND EMERGING ISSUES

Statistics Canada's population estimation system rests on two building blocks: (1) Census population counts, and; (2) components of population change, namely births, deaths and migrants. Postcensal estimates are carried forward by adding the components of population change over the subsequent years, to the base population, provided by the census. They are revised retrospectively when the next census counts become available. Thus, the census counts are both the base for the postcensal estimates, and the standard for their post-facto validation. The system has produced timely, reliable and internally consistent population estimates, and over the years has enjoyed a remarkable stability.

Much of its stability can be attributed to the high quality of the Canadian censuses. For Canada as a whole, undercoverage as measured by the Reverse Record Check (RRC) remained almost unchanged, at close to 2%, for three consecutive censuses - 1971, 1976 and 1981. Hence, even if the census fell somewhat short of the "true" population of Canada, it provided a highly reliable basis for gauging population growth.

The 1986 Census marks, however, a departure from the trend, as the rate of undercoverage, estimated by the Reverse Record Check, rose to 3.2%. The 1986 Census understates the population increase over the 1981-86 period by about 20%, if one accepts the component method as the standard of validation. Both the Reverse Record Check and the demographic analysis corroborate the deterioration of census coverage in 1986.

On the population components side of the equation - the other building blocks of the estimation system - records on births, deaths, and landed immigrants are fairly reliable. The interprovincial migration and emigration estimates have benefited from various data and

methodological enhancements, particularly since 1981, as was explained in the preceding section. But, as was also pointed out, they may suffer from various shortcomings inherent in any data sources – such as the family allowance and taxation files – that have been designed for administrative rather than statistical purposes. The estimates of interprovincial migration and emigration remain, along with census undercoverage and overcoverage, the prime sources of possible errors and biases in the postcensal estimates of population by province.

What does the future hold for the estimation system as described above? Can it continue working as it stands, or does it need some major reconceptualization? The apparently higher undercoverage rates of the 1986 Census, and its potential consequences for population estimates, has prompted the discussion of an alternative to the present census-based method of producing estimates. This alternative would no longer necessarily rely on the most recent census as a bench-mark, but instead would use relevant available information, including census counts, undercoverage and overcoverage, as well as administrative records, to generate the “best” possible estimates. In other words, the census counts remain an important ingredient of the estimation process, but not the overriding one; nor would the most recent census necessarily be used, if, say, the counts from the previous census were deemed to be more reliable.

After careful consideration, Statistics Canada has decided that the 1986 Census (unadjusted for undercoverage) would be used for the 1986 postcensal estimates and revision of the estimates for the 1981-86 intercensal period. In other words, the existing estimation procedures were reconfirmed. But at the same time, it was recognized that the evaluation of the census and estimates needed to be stepped up, and that an estimation strategy for the post-1991 Census period needed to be devised. Such an estimation strategy would have to take into account plans and realistic prospects for improvements and enhancements in the following four areas:

- (1) 1991 Census coverage;
- (2) Measurement of both undercoverage and overcoverage;
- (3) Administrative records used for the purpose of population statistics: enhancement of the currently used sources – Family Allowance and Taxation – and the harnessing of new ones, such as Old Age Security and Provincial Health Care Files;
- (4) Estimates of migration, particularly those concerning interprovincial migration, returning Canadian residents after a protracted stay abroad, and emigration from Canada.

These raise some fundamental issues concerning the philosophy and policy that ought to govern the working of a statistical system, thus transcending the rather narrow question of adjustment for undercoverage referred to at the outset of this paper. In the census-based conception, the emphasis is on the stability and internal coherence of the estimation system. In the conception of a census-divorced estimation model, a premium is placed on flexibility so as to increase the accuracy of the estimates through the utilization of the relevant available information, but possibly at the price of methodological consistency over time. The resolution of the dilemma between these two conceptions will be greatly influenced by the progress that is achieved in the four areas of statistical endeavour identified above.

ACKNOWLEDGEMENTS

The author has benefited from discussions with his colleagues in the Demography Division: Gwenaél Cartier, Gilbert Lagrange, Ronald Raby, Robert Riordan, Edward Shin and Ravi Verma. Valuable comments received from an anonymous referee and K.G. Basavarajappa,

David Binder, Malcolm Britton, Dick Carter, Ivan Fellegi and M.P. Singh, are also gratefully acknowledged.

REFERENCES

- CARTER, R.G. (1988). Measuring coverage errors in the census population. Presented at the annual meeting of the Canadian Population Society, the University of Windsor, Windsor, Ontario.
- COALE, A.J. (1955). The population of the United States in 1950 classified by age, sex and color. *Journal of the American Statistical Association*, 50, 16-54.
- FAY, R.E., PASSEL, J.S., ROBINSON, G.J., and CONRAN, C.D. (1988). The coverage of population in the 1980 Census. U.S. Department of Commerce, Bureau of the Census.
- FELLEGI, I.P. (1980). Should the census count be adjusted for allocation purposes - equity considerations. *Proceedings of the 1980 Conference on Census Undercount*, U.S. Bureau of the Census, 193-203.
- FREEDMAN, D.A., and NAVIDI, W.C. (1986). Regression models for adjusting the 1980 Census, *Statistical Science*, 1, 3-39.
- KEYFITZ, N. (1979). Information and allocation: Two uses of the 1980 Census, *The American Statistician*, 33, 45-50.
- KEYFITZ, N. (1980). Issues in adjusting for the 1980 Census undercount. Paper presented at the Annual Meeting of the American Statistical Association, Detroit.
- KISH, L. (1980). Diverse adjustments for missing data. *Proceedings of the 1980 Conference on Census Undercount*, U.S. Bureau of the Census, 193-203.
- LAPIERRE-ADAMCYK, E. (1970). Estimation of net census underenumeration by age and sex using demographic analysis techniques. Working paper, Demographic Analysis and Research Section, Statistics Canada.
- ROMANIUC, A., and RABY, R. (1980). The impact of census enumeration on selected Federal/Provincial transfer payments. Demography Division, Statistics Canada.
- SPENCER, B. (1980). Issues of accuracy and equity in adjusting for census undercoverage. Paper presented at the Annual Meeting of the American Statistical Association, Detroit.
- STATISTICS CANADA (1987). *Population estimation methods, Canada*. Catalogue No. 91-528E, Statistics Canada.
- STATISTICS CANADA (1988). Undercoverage rates from the 1986 Reverse Record Check. User Information Bulletin, No. 2, Ottawa.
- STOTO, M.A. (1987). Statement to the Subcommittee on Census and Population, Committee on Post Office and Civil Service, U.S. House of Representatives, San Francisco.
- WILK, M.B. (1981). Letter to Mr. H. Breau, Chairman, Parliamentary Task Force on Federal-Provincial Fiscal Arrangements, July 3, 1981.

Adjusting the 1986 Australian Census Count for Under-Enumeration

C.Y. CHOI, D.G. STEEL and T.J. SKINNER¹

ABSTRACT

In Australia, population estimates have been obtained from census counts, incorporating an adjustment for under-enumeration in 1976, 1981 and 1986. The adjustments are based on the results of a Post Enumeration Survey and demographic analysis. This paper describes the methods used and the results obtained in adjusting the 1986 census. The formal use of sex ratios as suggested by Wolter (1986) is examined as a possible improvement of the less formal use made of these ratios in adjusting census counts.

KEY WORDS: Census under-enumeration; Post-enumeration survey; Demographic estimates; Sex-ratios.

1. INTRODUCTION

The population census provides the basic information from which estimates are made of the population of the nation, each of the eight States and sub-State local government areas. In Australia, these population estimates are required for the determination of the number of seats each State will have in the Federal House of Representatives, the allocation of funds to each State, and the funding of local government authorities. Population estimates are also used in their own right as indicators of population growth and distribution and as denominators for various demographic, social and economic indicators. Because population estimates are used in such important ways, a high level of accuracy is required.

In Australia, it is known that the level of under-enumeration at the census is significant and that this level is related to important variables such as birthplace, geographic area and age/sex. Because of this, an adjustment for under-enumeration is made to census counts used for population estimates.

The adjustment of census counts for under-enumeration is a recent practice in Australia. Prior to the 1976 Census, census counts without adjustment for under-enumeration were used directly for population estimation purposes. The need to make this adjustment was recognised when the 1976 Census count fell considerably below the population estimates for the 1976 Census date which were updated from the 1971 Census, and when the 1976 Post Enumeration Survey (PES) showed a high under-enumeration rate of 2.6 per cent compared with 0.5 per cent in 1966 and 1.3 per cent in 1971. The 1976 PES also showed significant variations in under-enumeration between States and Territories, ranging from 4.2 per cent for the Northern Territory to 1.1 per cent for Tasmania. In 1986, the level of under-enumeration is estimated to be 1.9 per cent. As in 1976, there were significant variations between States and Territories. The adjustment of 1976 and subsequent census counts has been well received and no challenges have been raised to the appropriateness of doing so or the accuracy of the methods used. This is in contrast with the high level of controversy experienced in the United States of America on the appropriateness of making adjustments to the 1980 census counts for under-enumeration.

¹ C.Y. Choi, D.G. Steel and T.J. Skinner, Australian Bureau of Statistics, P.O. Box 10, Belconnen, ACT, 2616, Australia.

Data for the assessment of the level of under-enumeration are primarily derived from a census PES. Results of the PES are assessed by comparing these with estimates based on demographic statistics and other independent data such as statistics on school enrolments, on children whose parents receive government family allowances, and on persons registered with the government Medicare insurance system. In Australia, school enrolments for children aged 6-15 years are compulsory and until means-testing was introduced in November 1987, family allowances had been universally paid to mothers of all children of ages less than 17. Medicare insurance is also compulsory and universal for all residents. These independent statistics are therefore helpful as a check of the PES results and demographic estimates.

Although population estimates include an adjustment for under-enumeration, no adjustment is made for other census data. Census counts are published without adjustment.

2. THE 1986 POST-ENUMERATION SURVEY

In its five yearly population census, the Australian Bureau of Statistics (ABS) employs census collectors for the delivery of forms to each household and for the collection of completed forms from each household. The census is conducted on the basis of enumerating people where they are located on census night.

This collector-based field system allows the census collection phase to be completed two weeks after the census date. This allows a census PES to be conducted reasonably close to the census date – in 1986 within 4-5 weeks of census night. Because the PES asks a number of questions requiring detailed answers referring to a person's location on census night, its conduct close to census date minimises recall error and also reduces the number of exclusions due to deaths and overseas travel.

As the PES provides the basis for adjusting the census counts for under-enumeration, it is important that the PES be statistically independent of the census. The Appendix describes the steps taken to ensure independence.

The basic approach adopted in the 1986 PES was to select a sample of people independently of the census through a multi-stage area sample of private dwellings. The information required of each person in the selected households was obtained by personal interview of any responsible adult by trained field staff from the ABS regular interview panel. Matching of PES and census records to determine whether each person in the sample should have been included in the census and how many times the person was in fact included was undertaken by clerical staff employed in the Census Data Transcription Centre. The procedures used are described in the Appendix.

From the survey, the ratio of the number of persons who should have been included in the census (x) to the number of persons who were estimated to have been in fact included (y) can be estimated. This ratio is the net adjustment factor which accounts for both over and under-enumeration of individuals.

This adjustment factor, after weighting, is then applied to the actual census count (Y) to produce an estimate of the population (X), *i.e.* $X = Y (x/y)$.

To allow for differences in expected and actual sample take in the PES, this procedure was applied at the age (5 year groups), sex and geographic area (capital city statistical division/rest of State) level. PES estimates are produced on both an actual location at the census date and usual residence basis. The estimation also includes an adjustment for the small level of non-contact and non-response in the PES. For example the estimate of usual residence population for geographic area (s) and age sex cell (a) is:

$$X_{sa} = Y_{sa} x_{sa}/y_{sa}$$

where

$$x_{sa} = \sum_{gc} \frac{D_{gc} + d_{gc}}{D_{gc}} \cdot \frac{x_{sagc}}{f_g}$$

and

$$y_{sa} = \sum_{gc} \frac{D_{gc} + d_{gc}}{D_{gc}} \cdot \frac{y_{sagc}}{f_g} \cdot$$

In these estimation formulae the subscript *c* denotes the response status of the PES dwelling in the census and the subscript *g* denotes the geographic area in which the person was selected in the PES. *D_{gc}* is the number of responding dwellings and *d_{gc}* is the number of non-contact/non-responding dwellings in area *g* and census response category *c*. The sampling fraction varies between states and is denoted *f_g*.

In this form the estimator is a post-stratified ratio estimate. Ignoring for the moment that people may be enumerated in the census incorrectly or more than once, the estimator is the estimator obtained from a dual-record system or a capture-recapture approach discussed, for example, in Bishop, Fienberg and Holland (1975, pp231-234). This is shown in the diagram below where under the assumption of independence the estimate of the total population is *Y* (*x/y*) which is the ratio estimate *X*.

PES		
Census		
	Counted	Missed
	Counted	y
	Missed	Y
	x	

The 1986 PES, however, was *designed* to collect information on both the number of persons missed by the census and the number of persons over-enumerated, *i.e.* included in the census erroneously or included more than once. The estimate *X* takes into account both over and under-enumeration at the same time. In this respect, the approach adopted is different from the traditional capture-recapture methodology.

Variance estimation was based on treating *X* as a ratio estimate derived from a multi-stage sample. The relative standard errors on the PES estimates of the population are given in Table 1. From this table and tables 2 and 4 we see that standard errors are considerably less than the adjustments implied by the PES national age by sex estimates and State by sex estimates.

Table 1
1986 Census: Relative Standard Errors of PES Estimates
of the Population

Age	Males	Females	Persons
	%	%	%
0- 4	0.29	0.36	0.24
5- 9	0.29	0.30	0.22
10-14	0.28	0.29	0.21
15-19	0.32	0.32	0.24
20-24	0.49	0.43	0.34
25-29	0.49	0.36	0.32
30-34	0.39	0.34	0.27
35-39	0.36	0.30	0.24
40-44	0.38	0.32	0.26
45-49	0.37	0.30	0.25
50-54	0.43	0.38	0.30
55-59	0.38	0.30	0.25
60-64	0.41	0.38	0.29
65-69	0.43	0.37	0.29
70-74	0.53	0.41	0.34
75 +	0.47	0.39	0.31
All ages	0.12	0.10	0.08
State	Males	Females	Persons
	%	%	%
NSW	0.21	0.18	0.14
VIC	0.23	0.21	0.16
QLD	0.27	0.24	0.19
SA	0.27	0.20	0.17
WA	0.29	0.25	0.19
TAS	0.36	0.31	0.25
NT	1.65	1.53	1.22
ACT	0.61	0.74	0.55

For a more detailed description of the 1986 Post-Enumeration Survey and the estimation procedures, see Appendix.

3. DEMOGRAPHIC ESTIMATES OF CENSUS UNDER-ENUMERATION

An alternative method for the estimation of census under-enumeration is through the use of past demographic data including those from previous censuses, births and deaths registers, and overseas migration statistics. For example, estimates of the population at a certain date can be made by updating a previous census using data on births, deaths and overseas migration. The more distant is the previous census which serves as the base, the longer is the time series of reliable vital and migration statistics required, and the less reliance there needs to be on the accuracy of the census base. This is because estimates of persons born after the relevant census date will be affected only by the reliability of data on births, deaths and migration. Internal migration data in Australia are not sufficiently reliable to enable the use of demographic methods for estimating census under-enumeration at sub-national levels. Use of demographic estimates for census evaluation is therefore limited to Australian totals.

Australian data on births and deaths are available as a time series going back to the 19th century and it is unlikely that there have been significant omissions. Successive reports by the Australian Commonwealth Statistician after each population census from 1911 to 1961 claimed that the registration of births and deaths in Australia was substantially complete although it was recognised that some omissions were possible and that there were time lags in registrations. The Statistician's Report was discontinued after 1961. However, there is no evidence that the level of coverage of birth and death registrations has deteriorated since then.

Australia has also maintained comprehensive and reliable statistics on overseas arrivals and departures over a long period of time. These statistics cover all movements including permanent, long term and short-term movements. However, there are several deficiencies in the statistics on overseas arrivals and departures which limit their usefulness for the evaluation of the census data. First, there have been periods in the past when arrivals and departures were suspected of being inaccurately recorded (*e.g.* during World War II and the period immediately following the war). Second, because of the increase in overseas short-term movements since the 1960's only a sample (of about 1 in 20) of the arrivals and departure records has been processed for statistical purposes since 1971. Third, errors can occur in the classification of travellers into permanent, long-term and short-term categories. To avoid these errors of classifications the comparison of demographic estimates, census counts and PES estimates of the population at census date is made on the basis of actual location, which include all three categories of overseas movements.

For the assessment of under-enumeration at the 1986 Census, demographic estimates of the population as at census date 1986 by age and sex were made using births, deaths and overseas migration data going back to 1921 together with results of the 1921 Census. Demographic estimates of the population to age 65 years are therefore based solely on birth, deaths and migration data and would not be affected by the accuracy of the 1921 Census.

4. VALIDATION OF THE 1986 PES ESTIMATES

The following table shows the estimated population as at 30 June 1986 by age and sex based on demographic analysis and based on the 1986 PES. Medicare enrolments by age and sex are also shown.

There is a very high level of correspondence between PES and demographic estimates of the male population, particularly for those aged under 30. However, there is a large discrepancy for males aged 30-34, the demographic estimates being 20,000 higher than PES estimates. This can be attributed to a large net gain in the number of males of these ages from short-term movements into and out of Australia in the period 1981-86. Net gains from short-term movements of this magnitude are not detectable in the adjacent age-groups and therefore may reflect some error in overseas arrivals and departures statistics. With the volume of overseas movements being very high (over 6 million in 1986), a small error in reporting of age or in processing can lead to a relatively large discrepancy in the demographic estimate in net absolute term. The possibility of error in demographic estimates is further illustrated by the very high implied under-enumeration rate of 5.3 per cent for this age group compared with much lower rates for the surrounding age group.

It is, of course, quite likely that under-enumeration of overseas visitors was not adequately measured by the PES. However, in either case, errors in estimating the visitor component of the population should not affect the accuracy of official population estimates because these are based on the concept of usual residence and do not include visitors.

Table 2
Estimates of 1986 Population by Age and Sex Based on the 1986
PES and Demographic Analysis, and Medicare Enrolment

Age	Males ('000)									
	Population				Difference from Census			Percent Under-enumeration		
	Census (a)	PES(a)	DE(b)	Medi-care	PES	DE	Medi-care	PES	DE	Medi-care
0- 4	608.3	616.4	612.8	611.4	8.0	4.5	3.1	1.3	0.7	0.5
0- 5	594.9	602.4	603.0	612.3	7.5	8.1	17.4	1.2	1.3	2.8
10-14	660.8	670.4	668.4	674.3	9.6	7.6	13.5	1.4	1.1	2.0
15-19	673.1	688.4	687.7	693.1	15.3	14.6	20.0	2.2	2.1	2.9
20-24	648.5	679.5	681.3	681.3	31.0	32.8	32.8	4.6	4.8	4.8
25-29	649.2	677.7	675.0	688.5	28.5	25.8	39.3	4.2	3.8	5.7
30-34	615.5	630.0	650.1	647.5	14.5	34.6	32.0	2.3	5.3	4.9
35-39	622.2	634.2	632.7	646.2	12.0	10.5	24.0	1.9	1.7	3.7
40-44	504.2	512.6	517.0	522.3	8.4	12.8	18.1	1.6	2.5	3.5
45-49	419.8	427.0	416.5	436.8	7.2	-3.3	17.0	1.7	-0.8	3.9
50-54	363.7	371.2	371.4	377.9	7.5	7.7	14.2	2.0	2.1	3.8
55-59	373.4	379.5	384.9	386.6	6.1	11.5	13.2	1.6	3.0	3.4
60-64	341.1	347.0	348.1	350.6	5.9	7.0	9.5	1.7	2.0	2.7
65-69	259.6	263.6	251.8	265.5	4.0	-7.8	5.9	1.5	-3.1	2.2
70-74	204.2	208.2	200.8	213.0	4.0	-3.4	8.8	1.9	-1.7	4.1
75 +	229.5	233.0	181.5	250.1	3.5	-48.0	20.6	1.5	-26.4	8.2
Total	7768.3	7941.0	7883.1	8057.3	172.7	114.8	289.0	2.2	1.5	3.6

Age	Females ('000)									
	Population				Difference from Census			Percent Under-enumeration		
	Census (a)	PES(a)	DE(b)	Medi-care	PES	DE	Medi-care	PES	DE	Medi-care
0- 4	579.7	591.0	583.8	580.9	11.3	4.1	1.2	1.9	0.7	0.2
5- 9	565.1	572.4	565.5	582.1	7.3	0.4	17.0	1.3	0.1	2.9
10-14	628.0	636.8	630.2	641.8	8.8	2.2	13.8	1.4	0.3	2.2
15-19	644.1	657.4	651.4	666.3	13.3	7.3	22.2	2.0	1.1	3.3
20-24	633.1	652.5	644.4	670.4	19.4	11.3	37.3	3.0	1.8	5.6
25-29	648.7	660.7	665.4	684.1	12.0	16.7	35.4	1.8	2.5	5.2
30-34	618.1	627.8	631.2	643.9	9.7	13.1	25.8	1.5	2.1	4.0
35-39	612.1	619.1	600.2	626.3	7.0	-11.9	14.2	1.1	-2.0	2.3
40-44	482.6	488.6	489.6	495.4	6.0	7.0	12.8	1.2	1.4	2.6
45-49	399.1	403.0	397.9	411.6	3.9	-1.2	12.5	1.0	-0.3	3.0
50-54	349.1	354.6	343.9	358.6	5.5	-5.2	9.5	1.6	-1.5	2.6
55-59	362.6	366.5	362.4	372.4	3.9	-0.2	9.8	1.1	-0.1	2.6
60-64	358.2	364.4	351.3	365.3	6.2	-6.9	7.1	1.7	-2.0	1.9
65-69	298.2	302.2	301.9	306.7	4.0	3.7	8.5	1.3	1.2	2.8
70-74	259.0	262.9	262.2	269.7	3.9	3.2	10.7	1.5	1.2	4.0
75 +	396.2	404.7	385.0	434.1	8.5	-11.2	37.9	2.1	-2.9	8.7
Total	7833.8	7964.6	7866.2	8109.6	130.8	32.4	275.8	1.6	0.4	3.4

(a) Actual location basis.

(b) Demographic estimates based on 1921 Population Census and post 1921 demographic events.

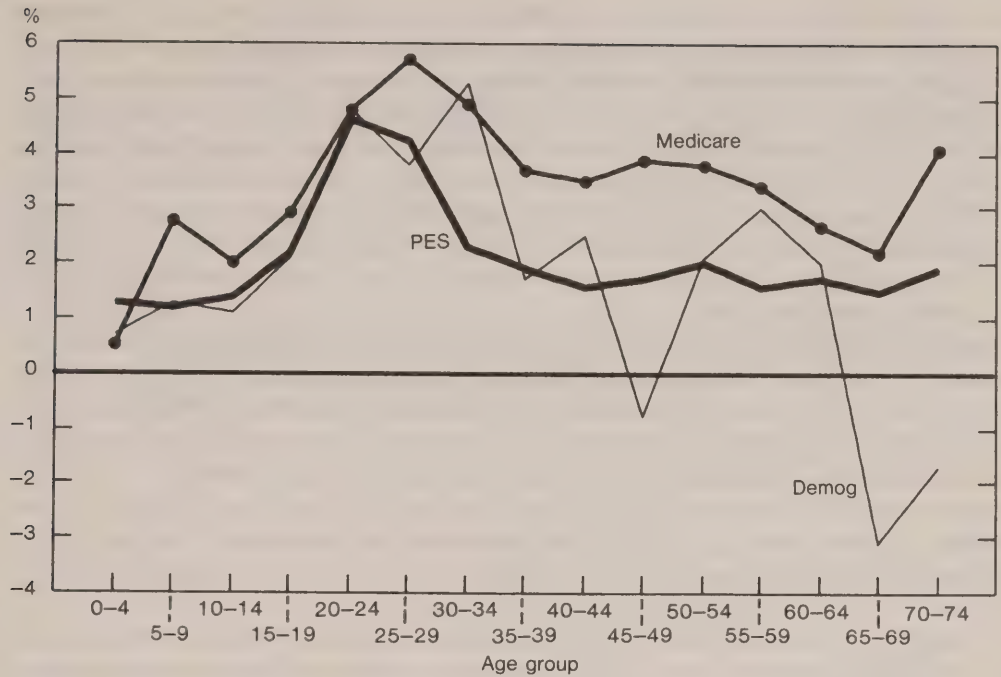


Figure 1. Percentage under-enumeration at the 1986 Census: Post-Enumeration Survey, Demographic Estimates and Medicare Enrolment-MALES.

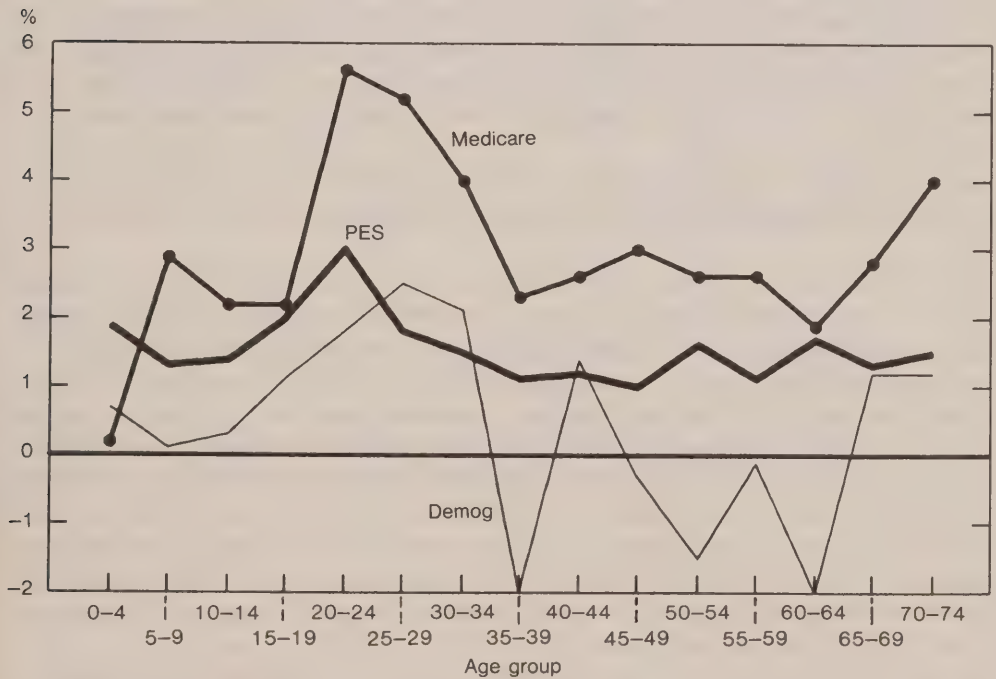


Figure 2. Percentage under-enumeration at the 1986 Census: Post-Enumeration Survey, Demographic Estimates and Medicare Enrolment-FEMALES.

For females, the level of correspondence between PES results and demographic estimates for ages below 35 is satisfactory. However, the demographic estimates for some age groups are considerably lower than PES estimates, and for those aged 35 to 39, and 45 to 64, they are lower than the unadjusted census count. Demographic estimates for these groups appear to be too low. This supports the view that demographic estimates are not sufficiently accurate for the production of population estimates and should be used only to assess PES results.

PES under-enumeration rates by age show a pattern which is smooth and much less erratic than that shown by demographic estimates. The higher PES rates for young adults aged 20-29 compared with those for other ages are as expected, given the higher rates of mobility among young adults, particularly males.

Medicare registrations are considerably higher than PES estimates and demographic estimates, except for the 0-4 age-group. Studies of registration practice show that the lower number in the 0-4 age group for medicare registration reflects the delays in births being registered with Medicare, and the higher numbers in other ages reflect delays in deleting from the Medicare register deaths and persons who have emigrated from the country.

Comparisons of PES estimates with estimates from family allowance registration and school enrolments for selected age-groups also show satisfactory correspondence. These results give some confirmation of the accuracy of the PES estimates in so far as the younger ages are concerned.

Although there is a satisfactory level of correspondence between PES estimates and other estimates of the population, there are two remaining problems which require consideration before the PES estimates can be accepted. The first emerges from an analysis of the PES estimates of census under-enumeration rates by age and sex. These rates are shown in Table 2.

Except for those aged 0-4 and 75 +, male under-enumeration rates are generally higher than female rates. While the rates for those aged 75 + could be affected by small sample size, the rate for females aged 0-4 appears too high, 1.9 per cent compared with 1.3 percent for males of the same age and for females aged 5-9. The number of females aged 0-4 estimated by the PES to have been under-enumerated was 11,300 compared with about 7,000 for the age group 5-9. This large difference in under-enumeration between those aged 0-4 and those aged 5-9 for females does not exist for males.

The PES sex ratio for persons aged 0-4 is 104.3 males to 100 female, lower than the census count ratio of 104.9 and the ratio of 105.0 males to 100 females estimated from demographic data.

On the above evidence, it appears that the PES has over-estimated females aged 0-4, although it is difficult to see how the PES could have over-estimated this group more so than other groups.

The second problem relates to the very high PES under-enumeration rate estimated for the Northern Territory. As shown in Table 4 it is 9.97% on an actual location basis and 6.45% on a usual residence basis. Northern Territory is a sparsely populated area (the census count in 1986 was 154,800 in an area of 1.3 million square kilometers) with a highly mobile population. The PES estimate of the population of Northern Territory is considerably higher than that based on the 1981 Census. Comparisons of PES estimates for the Northern Territory with independent estimates such as the number of children on the family allowance register and the number of school enrolments, also show that PES estimates are high. While these independent estimates may very well contain errors, it appears very likely that the PES has over-estimated the rate of under-enumeration for the NT.

The PES questionnaires were checked for the Northern Territory and were found to be satisfactory except for one collection district where problems with unreliable addresses and difficult terrain exposed inadequacies in field procedures and led to difficulties with matching.

Table 3
Comparison of 1986 PES Results with Independent Estimates

	PES Estimates	Demographic Estimates	Family Allowance	School Enrolment
Persons ('000)				
0- 4	1207.3	1196.6	1204.8(a)	-
5- 9	1174.8	1168.5	1177.0	-(b)
10-14	1307.2	1298.6	1304.2	1289.6

- (a) Family allowance registration for age 0 is understated because of the time lag in births being registered for family allowance. An adjustment was made by substituting the family allowance figure for age 0 by an estimate from the demographic analysis.
- (b) School enrolment not compulsory for children aged 5 years.

Table 4
PES Under-Enumeration Rates (%) by State

	Actual location basis	Usual Residence basis
New South Wales	1.54	1.51
Victoria	1.59	1.77
Queensland	2.68	2.43
South Australia	1.54	1.59
Western Australia	2.32	2.26
Tasmania	1.32	1.16
Northern Territory	9.97	6.45
Aust. Capital Territory	1.95	1.61
Australia	1.91	1.84

A judgement was made that the PES over-estimation of females aged 0-4 and of the NT population should be corrected by adjusting the PES results. The adjustment to females aged 0-4 was made by using the sex ratio from demographic estimates and applying this to the PES estimates of males aged 0-4. Essentially, this amounted to replacing the PES estimate of females aged 0-4 by a better estimate using the PES estimate of males and the sex ratio. The result of this adjustment was to reduce the estimates of this group by 4,000 to 587,000.

The problem with the NT estimates was handled by not using data from the problematic collection district. This reduced the Northern Territory under-enumeration rate to 9.1 per cent (on an actual location basis) and 5.5 per cent (on a usual residence basis).

The two adjustments to PES results reduced the overall national under-enumeration rate from 1.91 per cent to 1.87 per cent (on an actual location basis), or from 1.84 per cent to 1.81 per cent (on a usual residence basis). Table 5 shows PES estimates by age and sex after the above adjustments were made to the estimates for NT and for females aged 0-4.

Table 5
Census Count 1986 Adjusted for Under-enumeration by Age and Sex

Age	On the basis of 'actual location'					
	Males		Females		Persons	
	No. (^{'000})	% under enumeration	No. (^{'000})	% under enumeration	No. (^{'000})	% under enumeration
0- 4	616.3	1.30	586.6	1.17	1202.9	1.24
5- 9	602.4	1.24	572.4	1.27	1174.8	1.26
10-14	670.1	1.39	636.8	1.38	1306.9	1.39
15-19	688.3	2.19	657.3	2.02	1345.6	2.11
20-24	679.4	4.54	652.4	2.95	1331.8	3.76
25-29	677.5	4.17	660.7	1.81	1338.2	3.00
30-34	629.9	2.29	627.8	1.55	1257.7	1.92
35-39	634.0	1.87	618.9	1.11	1252.9	1.49
40-44	512.6	1.64	488.5	1.21	1001.1	1.43
45-49	426.9	1.66	403.0	0.98	829.9	1.33
50-54	371.2	2.04	354.6	1.56	725.8	1.80
55-59	379.5	1.62	366.5	1.06	746.0	1.34
60-64	347.0	1.70	364.4	1.70	711.4	1.70
65-69	263.6	1.52	302.3	1.35	565.9	1.43
70-74	208.2	1.92	262.9	1.47	471.1	1.67
75 +	233.0	1.49	404.7	2.08	637.7	1.86
All ages	7940.1	2.16	7959.7	1.58	15899.8	1.87
Age	On the basis of 'usual' residence					
	Males		Females		Persons	
	No. (^{'000})	% under enumeration	No. (^{'000})	% under enumeration	No. (^{'000})	% under enumeration
0- 4	615.3	1.29	585.9	1.22	1201.2	1.26
5- 9	601.3	1.23	571.2	1.22	1172.5	1.23
10-14	668.5	1.29	635.7	1.36	1304.2	1.33
15-19	685.6	2.11	654.3	1.97	1339.9	2.04
20-24	673.1	4.33	646.9	2.83	1320.0	3.59
25-29	672.6	4.02	657.2	1.80	1329.8	2.92
30-34	626.6	2.21	625.6	1.53	1252.2	1.87
35-39	630.9	1.78	616.7	1.05	1247.6	1.41
40-44	510.3	1.59	487.0	1.19	997.3	1.39
45-49	424.7	1.52	401.7	0.98	826.4	1.26
50-54	369.6	1.97	353.0	1.52	722.6	1.75
55-59	377.7	1.52	364.0	0.92	741.8	1.22
60-64	345.6	1.74	361.6	1.61	707.3	1.67
65-69	262.1	1.47	300.2	1.31	562.3	1.38
70-74	207.2	1.89	261.3	1.46	468.5	1.65
75 +	232.4	1.52	403.3	2.01	635.7	1.83
All ages	7903.6	2.08	7925.5	1.54	15829.1	1.81

5. ESTIMATING SUB-NATIONAL POPULATIONS

Internal migration data are not sufficiently reliable for demographic estimates of the population at sub-national levels to be used to assess census under-enumeration. However, a comparison of the 1986 PES estimates of the number of children aged 1-15 was made with the corresponding number receiving family allowance by State/Territory. This comparison shows a general agreement except for Northern Territory where the percentage difference was more than 2%.

Given this general agreement between PES estimates and family allowance data, and in the absence of reliable independent data on higher ages for comparison with PES estimates, the PES estimates (after adjustments) of the State and Territory populations were accepted.

Population estimates at the State/Territory level by age and sex, and at the local government area level were not derived directly from the PES. The 1986 PES was a sample survey and the results are subject to sampling error. Sampling errors at the State/Territory level by age and sex and at the local government area level are high, many unacceptably high, relative to the amounts of adjustment for under-enumeration which need to be made. An alternative indirect method, using an iterative proportional fitting (IPF) procedure, was used to produce State/Territory estimates by age and sex from those higher level PES estimates with a low sampling error. For a description of the IPF procedure, see Purcell and Kish (1979). This procedure involved taking the national population estimates by age and sex and the State/Territory estimates within each sex and adjusting the census age by State/Territory counts to these two margins.

The IPF procedures involves the following cycles $n = 0, 1, \dots$

$$X_{gas}^{(2n+1)} = X_{gas}^{(2n)} \frac{X_{as}}{X_{as}^{(2n)}} (2n)$$

$$X_{gas}^{(2n+2)} = X_{gas}^{(2n+1)} \frac{X_{gs}}{X_{gs}^{(2n+1)}}$$

and $X_{gas}^{(0)} = Y_{gas}$ the census count for state g , age category a and sex s . The procedure converges to a unique solution. The use of IPF procedures, of course, assumes that the relationship between the variables within the association structure is valid and that this relationship is preserved.

For estimates for local government areas, the problem with high sampling error is more acute and results of the PES are not sufficiently reliable to make direct estimates of under-enumeration for each local government area. Based on the premise that under-enumeration is age/sex and birthplace (Australian born/Overseas born) selective, and that it differs between States/Territories and between capital city and the rest of the State, adjustments for under-enumeration at the local government area level were made to reflect under-enumeration differentials by age, sex, capital city/rest of State and Australian-born/overseas-born.

6. PROBLEMS WITH THE PES ESTIMATION

As pointed out by Bailer (1985), for example, the bias and consistency of the PES estimates is affected by errors in the matching process, any correlation between a person being missed in the census and in the PES, and erroneous inclusions in either the census or the PES. It is

because of the possible effects of these factors that the results of the PES are assessed using demographic and administrative data in the ways described above.

Errors in matching will bias the PES estimates. Failure to match records that in fact should match will lead to the creation of apparently under-enumerated persons and the PES estimate will be an over estimate. The effect of false matches will be the reverse.

Erroneous inclusions in either the census or PES will inflate the values of Y or x and hence the PES estimate. The US Bureau of the Census conducts a special "E-sample" selected from the census to estimate the extent of erroneous inclusions in the census which can then be incorporated in the estimate by adjusting the census count Y . For a description of the E sample, see Fay, Passel and Robinson (1988). The matching and estimation procedures used by the ABS attempt to adjust for some of the effect of erroneous inclusions by determining not only whether or not someone has been included but whether they should have been included and if they have been included more than once. For example in the 1986 PES, 250 people were determined to have been included twice and four persons had been included three times. Cases were also found where persons had been included but should not have been. In this way viewing the PES estimation as a ratio estimator rather than a dual system estimator enables the accounting for some erroneous inclusions.

The dual system estimation method makes the assumption that whether or not someone is missed in the PES is independent of whether or not that person is missed in the census. Whilst all practical steps have been taken in ensuring that the two field and processing systems involved in the collections are completely separate and independent it is still possible for correlation to exist. Positive correlation will mean that the PES estimate based on the assumption of independence will be an under-estimate, negative correlation leads the PES estimate to over-estimate. Negative correlation would occur if being included in the census led people to be hard to enumerate in the PES but we have no clear evidence for this; the final response rate for the PES (95%) is in line with other household surveys conducted by the ABS. Positive correlation seems more likely, and there appears to have been some evidence of this in the 1981 Census. If such positive correlation exists then the PES based adjustments will have not gone far enough but will have been in the right direction.

7. ALTERNATIVE METHODS OF ESTIMATION (WOLTER 1986)

The idea of combining PES data and demographically derived sex ratios or sex ratios obtained from other sources is the basis of methods suggested by Wolter (1986). Wolter suggests several models and associated methods which formally combine sex ratios and PES estimates. These methods are attempts to loosen the assumption of independence inherent in the PES estimation methods.

Wolter considers two models. In the first it is assumed that the degree of association in under-enumeration between the PES and the census (as measured by the cross-product ratios in tables such as the diagram shown earlier in this paper) is the same for males and females within each age category. In the second model independence is assumed for females and an externally derived sex ratio is used to obtain the male figure. It is then possible to calculate the cross-product ratios implied for males.

From an initial evaluation of these methods applied to Australian data, it was found that the first model produced very erratic estimates of the cross product ratios, with approximately 50% being negative. This was greatly reduced under the second model although some remained negative and were set to zero in a modified model. The problem with negative cross-product

Table 6
Sex Ratios: Males per 100 Females

Age	Alternative	PES
0- 4	105.0	104.3
5- 9	105.2	105.2
10-14	105.2	105.3
15-19	104.7	104.7
20-24	104.1	104.1
25-29	102.6	102.6
30-34	100.3	100.3
35-39	102.4	102.4
40-44	104.5	104.9
45-49	105.2	106.0
50-54	104.2	104.7
55-59	103.0	103.5
60-64	95.2	95.2
65-69	87.1	87.2
70-74	78.8	79.2
75 +	57.9	57.6

ratios was also identified by Wolter (1986, p. 7). The second model, modified, was then applied to 1986 data. For age groups 5-9 up to 35-39, the sex ratio obtained from the PES were in line with expectations and those sex ratios were used giving exactly the PES estimate. For the 0-4 age group the sex ratio obtained from demographic estimates was used and for the 40-44 to 75 + age groups, an alternative estimate of the sex ratios based on census counts was used. The sex ratios are given in Table 6.

The sex ratio used and the PES sex ratios are not greatly different so applying Wolter's second model leads to only small changes in the PES estimates. For the 0-4 and 75 + age groups the estimates of males are increased by 0.7% and 0.5% respectively. For the 45-49 and 70-74 age groups the estimates are reduced by between 0.7% and 0.5%. This analysis suggests that the differences in biases between sexes in the PES estimation method due to the combined effect of the potential problems discussed above, are relatively small. It could be the case that any biases are affecting males and females to an approximately equal degree so that PES sex ratios are broadly acceptable.

Our experience in 1981 and 1986 demonstrated the need to use sex ratios in assessing measures of under-enumeration and we believe the Wolter method is a useful way of generating alternative estimates against which the Census count and direct PES estimates can be judged. The general acceptability of the PES sex ratios in 1986 has meant that using this method made little difference. The acceptability of the PES sex ratios in 1986, except for the 0-4 age group contrasts with the experience in 1981, where an adjustment to the PES estimates was considered necessary for a number of age groups based on alternative sex ratios. These differences in the 1981 and 1986 experience may reflect a reduction in correlation between under-enumeration in the census and the PES in 1986.

8. CONCLUSION

While the ABS has adjusted the past three censuses for under-enumeration, our confidence in the basic reliability of the PES stems from its general consistency with other data sources. No fundamental change in approach is anticipated for the next census to be conducted in 1991. However, we believe there is a need to investigate further potential causes of bias, in particular the adequacy of the clerical matching procedures, and methods to overcome correlation bias. It is also planned to investigate the possibility of creating a demographic data bank on a usual residence basis, so that the effects of the large volume of short-term movements can be eliminated or reduced.

ACKNOWLEDGEMENT

We are grateful to the reviewers of this article for their comments and suggestions for further research.

APPENDIX

THE 1986 POST-ENUMERATION SURVEY

General

The 1986 PES was conducted in the 4th and 5th weeks after census night. The survey involved interviews with a sample of the population from about 35,000 private dwellings (2/3 of one percent of dwellings) across Australia involving about 100,000 persons. The sampling fraction varied between States and Territories, with the smaller States and Territories having higher sampling fractions. Personal data on name, age, sex, marital status and birthplace were obtained by interviewers for matching with information on the census form. For each person in the survey, information was sought on their place of usual residence, where they spent census night, their address before and after census night and any other address where they might have been included on a census form. At each given address, the personal information was matched to census forms to establish whether a person was missed, counted once or the number of times counted if counted more than once.

Scope and Sample Structure of the PES

Except for the special cases mentioned below, the PES included in its scope all persons who should have been enumerated in the census, except those who had gone overseas or died between the census and PES dates. Diplomatic representatives and persons in diplomatic dwellings were not included in the census. These persons were excluded from the survey as were babies born after census night. Persons in the survey who were overseas on census night were matched to census forms to determine whether they were incorrectly included in the census.

For practical reasons, very sparsely settled areas were not included in the PES. In these areas, special census procedures were used to contact and enumerate Aboriginal groups, people in mining camps, cattle stations, etc. The PES in these areas would need to rely on the same contacts and procedures adopted for the census and therefore could not accurately and independently measure under-enumeration. Consequently, the scope of the PES excluded these areas.

Non-private or special dwellings such as hospitals, hotels, and motels also were not included in the PES. The vast majority of residents in non-private or special dwellings would have been short-term residents and, according to normal ABS survey rules short-term residents would have a chance of being included in the survey at their place of usual residence where information on such persons would be obtained. A relatively small number of long term residents of these dwellings were consequently not included in the PES. For estimation purposes, populations out-of-scope were assumed to have the average capital or non-capital city rate of under-enumeration for each State as appropriate and the average Territory rate for each of the two Territories.

As non-private or special dwellings and sparsely settled areas contained less than 3% of the total population, any differences in under-enumeration of these areas compared with areas covered by the PES would be unlikely to have a significant effect on the overall estimated level of underenumeration at the State or National level.

Interaction Between the Census and the PES

It is important that the PES be conducted as independently of the census as possible. Otherwise, the factors that led to a person being missed or overcounted in the census may also be present in the PES, resulting in biased estimation of the under-enumeration. Furthermore, knowledge of the areas to be included in the PES might influence the performance of census collectors in these areas so that the PES sample would not be a representative sample of the under-enumeration. For these reasons the field and office staff used in the census and PES were totally separate. PES interviewers were not employed as census collectors or census group leaders, and census field staff were not told which areas were included in the PES.

Independence was further guaranteed in two ways – by ensuring the operational independence of the field systems, and by adopting special procedures for census forms received by mail after the PES field work commenced.

To ensure operational independence, PES field work commenced after all available census forms had been collected from the field. Thus census collectors were not in the field at the same time as PES interviewers and there was no possibility of interaction, even unintentional, between census and PES field staff.

Special procedures for census forms received after the PES commenced were required to overcome the effects PES fieldwork may have had on householders who were late returning their census forms. In some cases, PES interviewers discovered census forms still uncollected. This situation was possible because some people had preferred to post in their census forms and had not yet done so, or the census collector had been unable to make contact to collect them. Some of these people who were included in the PES may have been prompted to post their forms in, where they would not otherwise have done so. To overcome this potential bias, any census form returned by mail after Monday 20 July 1986 (the day PES interviewing commenced) was considered a late form. Special procedures for the treatment of late forms are described later in this Appendix.

Matching procedures of the PES

Matching for the purpose of determining whether a person was missed, counted once or the number of times counted if counted more than once, was conducted in two stages. Both these stages were clerical processes undertaken by staff at the census Data Transcription Centre.

The first stage was the locating of census forms for the addresses of households selected in the PES. Processing of 1986 Census forms were centralized in Sydney. Staff at the Population Census Data Transcription Centre were requested to compare the address on the front of the PES interview form with all addresses given in the record book of the census collector

who was responsible for the collection district (CD) in which the PES household was located. The record book was used as a control in the delivery and collection of census forms, and contained information such as name, address and number of persons for all households in the CD.

To assist identification of households where addresses were sometimes vague, for example in rural areas, processing staff were asked to also use names of the householders, property names etc. In addition, staff were instructed to check through all addresses in the record book so that any duplicate census forms were identified. Addresses in record books of adjacent CDs were also checked if the address of the household selected by the PES was near the boundary of the CD.

The second stage was person-matching and this was based on the name and demographic details of the persons listed on the census and PES forms. In this matching process, a search form was generated for each address reported in the PES for any person in the household, other than the address of the PES selected dwelling. A search form was treated the same way as a PES interview form and an attempt was made to locate the census form which corresponded to the search form address.

In most cases, the person-matching procedure was straight forward. There were, however, cases of spelling errors and insufficient details on addresses to identify a clear match on name. In these cases, a judgement on whether or not a person was counted was made based on other information such as age, sex, marital status, birthplace and relationship to other members in the census household. For doubtful cases, processing staff were required to consult their supervisor.

The PES also asked the respondent whether each person was included on a census form. When matching failed because of lack of adequate information, the respondent's statement about whether or not the person was counted was accepted. There were a few cases where even this information was unavailable. These cases were considered not counted in the census.

After matching, the data was entered onto computer tapes, edited and reformatted to produce a clean unit record file giving the number of times person in the PES sample were counted in the census.

Treatment of Late Census Forms and 'Dummy' Census Forms

In forming the estimation equation:

$X = Y (x/y)$, where

X = estimated census count adjusted for underenumeration

Y = raw census count, unadjusted

x = PES estimate of the number of persons who should have been included in the census and

y = PES estimate of the number of persons who were included in the census,

two categories of census forms were treated as missed in the census. These are 'dummy' census forms and late census forms.

Dummy census forms were created during census fieldwork for dwellings at which households were known to be residing, did not return their census forms and could not be contacted. Census collectors were instructed to exercise extreme care in creating these dummy forms and they needed to be satisfied that there was concrete evidence that the dwellings were occupied on census night. The collectors were instructed also to obtain as much information as possible regarding the number and the demographic characteristics of these residents.

When a PES address was matched to a dummy census form, the lack of name and reliable personal characteristics on the census form made it impossible to perform the matching operation satisfactorily.

It is also necessary to handle late census forms differently from normal census forms. Because late census forms might have been prompted by a PES interviewer calling, their inclusion could lead to a bias in the estimation of under-enumeration.

In the 1986 Census, there were 115,000 persons recorded on dummy census forms or late census returns, or 0.7 per cent of the population. Both dummy and late census forms were excluded from the raw census count (Y) and the PES estimate of the number of persons who were counted in the census (y), but were included in the PES estimate of the number of persons who should have been counted in the census (x). In other words, persons on dummy and late forms were treated as missed and adjusted for by (x). The adjustment factor (x/y) is exaggerated because of the exclusion of dummy and late forms from (y), but this exaggeration is compensated for by the exclusion of these forms from the raw census count (Y).

Estimation Procedure

The estimation procedure was applied at the age by sex by geographic area (capital city statistical division/rest of state) level. Adjustment factors were included in the estimation formulae to partly account for non-responding and non-contact households. These factors adjust both of the main estimates, x and y , by effectively imputing, for each non-contact or refusing household, the average number of persons per household, and, for each person so imputed, the average rate of under-enumeration at the relevant age by sex by area level. To reduce the bias from the use of such adjustment factors, the factors were calculated for various subgroups of households by the status of enumeration at the census (such as occupied dwelling, late returned form). This enumeration status was considered to be related to what non-response was encountered in the PES.

REFERENCES

- BAILAR, B.A. (1985). Comments on "Estimating the population in a Census Year: 1980 and beyond" by E.P. Ericksen and J.B. Kadane, *Journal of the American Statistical Association*, 80, 109-114.
- BISHOP, Y.M.M., FIENBERG, S.A., and HOLLAND, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge: MIT Press.
- FAY, R., PASSEL, J.S., and ROBINSON, J.G. (1988). The coverage of population in the 1980 Census. Evaluation and Research Report, PHC 80-E4, United States Bureau of the Census, Washington D.C.
- PURCELL, N.J., and KISH, L. (1979). Estimation for small domains. *Biometrics*, 35, 365-384.
- WOLTER, K.M. (1986). Capture-recapture estimation in the presence of known sex ratio. SRD Research Report, United States Bureau of the Census, Washington, D.C.

When Are Census Counts Improved by Adjustment?

NOEL CRESSIE¹

ABSTRACT

There are persuasive arguments for and against adjustment of the U.S. decennial census counts, although many of them are based on political rather than technical considerations. The decision whether or not to adjust depends crucially on the method of adjustment. Moreover, should adjustment take place using say a synthetic-based or a regression-based method, at which level should this occur and how should aggregation and disaggregation proceed? In order to answer these questions sensibly, a model of undercount errors is needed which is "level-consistent" in the sense that it is preserved for areas at the national, state, county, *etc.* level. Such a model is proposed in this article; like subareas are identified with strata such that within a stratum the subareas' adjustment factors have a common stratum mean and have variances inversely proportional to their census counts. By taking into account sampling of the areas (*e.g.*, by dual-system estimation), empirical Bayes estimators that combine information from the stratum average and the sample value, can be constructed. These estimators are evaluated at the state level (51 states, including Washington, D.C.), and stratified on race/ethnicity (3 strata) using data from the 1980 post-enumeration survey (PEP 3-8, for the noninstitutional population).

KEY WORDS: Empirical Bayes estimation; Loss functions; Measures of improvement; Quantile function; Spatial correlation; Synthetic estimation.

1. INTRODUCTION

This article is of a technical nature, but it is important to present a brief explanation of the political and social ramifications of the "undercount issue" in the United States of America. By December 31 of the year of the decennial census, the U.S. Census Bureau is specified by law to submit state population counts to Congress for the purpose of reapportionment of the House of Representatives, and by March 31, 1991, to submit small-area population counts for the purpose of redistricting. In recent decades, the number of uses to which census data are put have multiplied: revenue-sharing formulas use population and per capita income for each incorporated place, demographic and sociological research at regional, state, and national levels usually rely on census counts, *etc.*

Inaccurate census counts should be cause for concern to the whole nation. That certain groups of people (young black males, illegal aliens, *etc.*) are harder to count than others, is without question; see Ericksen and Kadane (1985), and Freedman and Navidi (1986), and the discussion following these articles. If the hard-to-count groups were distributed in equal proportions throughout the political and administrative regions of the USA there would be far less controversy over what to do about the uncounted people. As it is, many of the large American cities such as Chicago, Detroit, New York, and Los Angeles feel they are losing federal funds because their cities contain more of the types of people that tend to remain uncounted. And certain states such as New York and California feel they are under-represented in Congress, to the benefit of Midwestern states such as Indiana and Iowa.

¹ Noel Cressie, Department of Statistics, Iowa State University, Ames, IA 50011.

Census undercount is defined simply as the difference between the true count and the census count, expressed as a percentage of the true count. My approach to its estimation is model-based, relying on data obtained from the post-enumeration survey (PES). A number of technical aspects of a model-based approach to adjustment will be addressed in this article. Section 2 establishes the model, addresses the question of choice of measures of improvement, and presents results for aggregation and disaggregation based on Bayes and Synthetic estimators. Section 3 gives *empirical* Bayes versions of the results of Section 2. Section 4 summarizes what has been learned from this model-based approach; there is also discussion of the implications of the sufficient conditions that guarantee risks of adjusted counts to be smaller than risks of census counts.

2. THE MIXTURE MODEL AND ITS CONSEQUENCES

At the outset I would like to explain the source of random variation in my model, originally defined in Cressie (1986), and further developed in Cressie (1988). I consider the true population in any well-defined stratum of the USA, to be unknown. After observing the corresponding census population, the uncertainties about the true population are updated. In other words, all inference will be performed *conditionally* on the observed census counts.

2.1 The Model

The method of *synthetic estimation* constructs estimators of undercount at a particular level (e.g., the state level) by summing undercounts of various strata (e.g., demographic strata) over the area being considered (e.g., California), where it is assumed that any stratum has a *constant* proportion of true counts to census count regardless of which area is being considered. For example, it would be assumed that the proportion for young black males is the same for California, Delaware and so on. Most often these strata are defined demographically according to the factors of age, race, and sex. However Tukey (1981) suggested that geographic and urban factors should be added. Two such stratifications of the USA are given in Isaki *et al.* (1986).

The mixture model I am proposing assumes a stratification has been defined already, although in Section 4 there is a suggestion how one might determine *post hoc* whether a chosen stratification is satisfactory.

Suppose there are $j = 1, \dots, J$ strata, and $i = 1, \dots, I$ areas (e.g., at the enumeration-district level, $I \approx 300,000$, while at the state level, $I = 51$, including the District of Columbia; for demographic stratification, $J = 30$ say, while for the two stratifications in Isaki *et al.*, 1986, $J = 90$ and $J = 96$. Think of stratum j as fixed (for example, stratum j might be the blacks in central cities in those SMSA's whose population's greater than or equal to 250,000, in the New England Census Division). Then as i ranges from $1, \dots, I$, a sequence of subareas is generated; the subarea indexed by " ji " refers to that part of the i -th area that has stratum j in it. Only subareas with *nonzero census counts* are considered.

Define

$Y_{ji} \equiv$ true count in the j -th stratum of area i (2.1)

$C_{ji} \equiv$ census counts in the j -th stratum of area i (2.2)

$F_{ji} \equiv Y_{ji}/C_{ji}; i = 1, \dots, I; j = 1, \dots, J.$ (2.3)

Suppose for the moment that we know the ratios $\{F_{ji}: j = 1, \dots, J\}$ for the i -th area. Then from the census counts C_{ji} , the true count Y_i can be calculated.

$$Y_i = \sum_{j=1}^J F_{ji} C_{ji}. \quad (2.4)$$

The F_{ji} are often called *adjustment factors*. The strata are constructed so that these adjustment factors $\{F_{ji}: i = 1, \dots, I\}$ are as homogeneous as possible within the j -th stratum; $j = 1, \dots, J$ (Tukey 1981).

Realistically the adjustment factors are never known; synthetic estimators exploit the homogeneity and replace (2.4) with

$$Y_i^{\text{syn}} = \sum_{j=1}^J F_j C_{ji}. \quad (2.5)$$

Now there are only J synthetic adjustment factors $\{F_j: j = 1, \dots, J\}$ to estimate, which through (2.5) yields an estimate of Y_i . Synthetic estimators have the advantage that the adjustment factors are independent of i and so can be applied to *any* level of aggregation.

The (estimated) adjustment factors could also be modeled by regression on independent variables that may or may not be census variables; for example, percent minority, crime rate, and percent conventionally counted in the census. Consider,

$$Y_i^{\text{reg}} = \sum_{j=1}^J \left(\sum_{k=1}^p \beta_{k,j} z_{k,ji} \right) C_{ji}. \quad (2.6)$$

To fit the parameters $\beta_{1,j}, \dots, \beta_{p,j}$ efficiently, various assumptions are made about the error components $\{F_{ji} - \sum_{k=1}^p \beta_{k,j} z_{k,ji}\}$, viz. independent and identically distributed with mean zero.

Ericksen and Kadane (1985) propose the fitting of a regression relation to $\sum_{j=1}^J F_{ji} C_{ji} / \sum_{j=1}^J C_{ji}$; $i = 1, \dots, I$. Freedman and Navidi (1986) criticize the approach and point out the consequences of failure of any of the error assumptions. A problem they did not perceive which I emphasize in (2.7) below, is the heteroskedasticity forced onto the problem by working with ratios; Section 2.2 justifies this model choice. Furthermore, in this latter regression approach undercounts across strata are combined, so that variation between strata is shared by both the regression relation and the error variance. More precise estimators can be obtained through (2.6) by allowing each stratum its own regression relation. Homoskedastic errors and a regression model based on the combination of heterogeneous strata, are also assumed by Ericksen and Kadane (1987) and Ericksen, Kadane and Tukey (1987). It seems that the combination of heterogeneous strata was made necessary by the lack of suitable data.

I do not assume F_{ji} 's that depend only on j , nor a regression relation for the F_{ji} 's, but instead reformulate the synthetic assumption $F_{ji} \equiv F_j$, into a (statistical) homogeneity assumption:

$$F_{ji} \sim N(F_j, \tau_j^2 / C_{ji}); \quad i = 1, \dots, I; \quad j = 1, \dots, J, \quad (2.7)$$

where " \sim " means "is distributed as," and $N(\mu, \sigma^2)$ is a normal distribution with mean μ and variance σ^2 . Using a regression relation for the mean has the potential of explaining more of the variation of the F_{ji} 's at the risk of introducing bias through misspecification. The strata chosen in Section 3 are based on race; it was decided not to cloud this sensitive issue with selection of controversial regression variables. I shall refer to the model (2.7) as a mixing

distribution. *The normality assumption is made for convenience and will be relaxed later.* Here F_j is a fixed but unknown mean to be estimated, and $\tau_j^2 = \text{var}(\sqrt{C_{ji}} F_{ji})$ is a parameter I shall call the (standardized) *stratum variance*. As a representation of reality, model (2.7) is better at higher levels of aggregation; see Section 3. All distributions in (2.7) are assumed independent.

There are good reasons for weighting the variance by $1/C_{ji}$ (see Cressie 1987a, Appendix and 1988). The most attractive consequence of model (2.7), is that it is *level-consistent*; that is, it is preserved through different levels of aggregation. Specifically,

$$F_{j,i\&i'} \sim N \left(F_j, \frac{\tau_j^2}{C_{j,i\&i'}} \right), \quad (2.8)$$

where

$$F_{j,i\&i'} \equiv \frac{F_{ji}C_{ji} + F_{ji'}C_{ji'}}{C_{j,i\&i'}}, \text{ and } C_{j,i\&i'} \equiv C_{ji} + C_{ji'}. \quad (2.9)$$

This is a very important property that most of the currently proposed statistical models of undercount do *not* possess. It enables the modeler to escape from the geographical and historical accidents that divided up the country into the states, counties, *etc.*, that we now see.

Of course the $\{F_{ji}: i = 1, \dots, I; j = 1, \dots, J\}$ are not available as data; if they were, $\{Y_i: i = 1, \dots, I\}$ would be trivial to calculate. In reality, some sampling takes place so that F_{ji} is observed imperfectly. The best way to think of it is that within stratum j of the i -th area, a sample is taken for undercount. Let the outcome be X_{ji} (e.g., X_{ji} is the ratio of dual-system estimator to census count, for the j -th stratum in the i -th area), and model

$$X_{ji} \sim N(F_{ji}, \sigma_j^2 / C_{ji}); i = 1, \dots, I; j = 1, \dots, J, \quad (2.10)$$

where F_{ji} is an unknown mean parameter to be estimated, and $\sigma_j^2 = \text{var}(\sqrt{C_{ji}} X_{ji})$ is a parameter I shall call the (standardized) *sampling variance*. All distributions in (2.10) are assumed independent. When the number of strata is large, a large PES (say, 300,000 households) is needed to obtain data for each area-stratum combination.

Probability-proportional-to-size sampling was used by the U.S. Census Bureau in its 1980 post-enumeration program, which implies a sampling variance of the form given in (2.10). As a consequence of this weighting, (2.10) is also level-consistent.

2.2 Loss Functions (Measures of Improvement) and their Bayes Estimators

The term loss function is used in statistical decision theory (see, for example, Ferguson 1967) to quantify the loss incurred from using $\hat{\theta}$ as a parameter estimator when the true value is θ . For example, a squared-error loss function is $(\hat{\theta} - \theta)^2$. Adopting a more optimistic terminology, the Census Bureau decided in 1986 to use "measure of improvement" instead of "loss function."

Think of (2.10) as a conditional distribution of X_{ji} given F_{ji} , and (2.7) as the mixing (or "prior") distribution of F_{ji} . To predict F_{ji} then, the "*posterior*" distribution of F_{ji} given X_{ji} is needed. Notice that a Bayesian terminology is being used since I am thinking of the F_{ji} as random variables whose collection is modeled according to (2.7). But as well as these random parameters, there are fixed but unknown parameters $\{F_j\}$, $\{\tau_j^2\}$, $\{\sigma_j^2\}$ to be estimated. The posterior of $F_{ji} | X_{ji}$ is,

$$\frac{(\text{distribution of } X_{ji} \mid F_{ji}) \cdot (\text{“prior” of } F_{ji})}{\text{marginal of } X_{ji}} \tag{2.11}$$

For squared-error loss, the usual Bayes estimator of F_{ji} is simply the expectation of F_{ji} with respect to the posterior: $F_{ji}^{uba} = E(F_{ji} \mid X_{ji})$. Substituting the model (2.7), (2.10) into (2.11), the posterior distribution is easily obtained (see, for example, Lindley and Smith 1972):

$$F_{ji} \mid X_{ji} \sim N\left(F_j + \frac{\tau_j^2}{\tau_j^2 + \sigma_j^2} (X_{ji} - F_j), \frac{\sigma_j^2 \tau_j^2}{\tau_j^2 + \sigma_j^2} / C_{ji}\right), \tag{2.12}$$

for $i = 1, \dots, I; j = 1, \dots, J$. Hence the posterior expectation is simply

$$F_{ji}^{uba} = F_j + D_j (X_{ji} - F_j), \tag{2.13}$$

where $D_j \equiv \tau_j^2 / (\tau_j^2 + \sigma_j^2)$. To convert (2.13) into an empirical Bayes estimator, estimators have to be found for F_j and D_j ; see Section 3.1.

Although the normality assumptions in (2.7) and (2.10) were used to derive (2.13), more generally (2.13) can be shown to be Bayes for squared-error loss, when assuming simply the mean and variance structure of (2.7) and (2.10), and $E(F_{ji} \mid X_{ji}) = a_{ji} + b_{ji}X_{ji}$. Goldstein (1975) has an even more general result of which this is a special case. For ease of exposition I shall continue to assume normality but it should be remembered that there is a nonparametric optimality for all the estimators considered.

The estimator F_{ji}^{uba} given by (2.13) is Bayes for squared-error loss, within the j -th stratum of the i -th area. Define the estimator of Y_i ,

$$Y_i^{uba} \equiv \sum_{j=1}^J F_{ji}^{uba} C_{ji}; i = 1, \dots, I, \tag{2.14}$$

and consider the following general loss function:

$$\sum_{i=1}^I (Y_i^{est} - Y_i)^2 f(C_i), \tag{2.15}$$

where $f(C_i)$ is any nonnegative function of the i -th area's census count. Minimizing (2.15) over all $Y_i^{est} \equiv \sum_{j=1}^J F_{ji}^{est} C_{ji}$ leads to choosing F_{ji}^{est} 's such that $E[\sum_{i=1}^I \sum_{j=1}^J \lambda_{ji}^{est} (F_{ji}^{est} - F_{ji})^2 \mid \{X_{ji}: i = 1, \dots, I; j = 1, \dots, J\}]$ is minimized, where the $\lambda_{ji} \geq 0$ only depend on census counts $\{C_{ji}: i = 1, \dots, I; j = 1, \dots, J\}$. This minimum is achieved by the estimator (2.14), which shows it to possess a certain robustness since it is optimal regardless of which $f(\cdot)$ is chosen.

In accordance with recommendation 7.2 in National Academy of Sciences (1985), choice of $f(C_i) = 1/C_i$ yields an area's contribution to the total loss that reflects the size of its population. Among the loss functions the Census Bureau has been using, the one most like (2.15) with $f(C_i) = 1/C_i$, is

$$\sum_{i=1}^I (Y_i^{est} - Y_i)^2 / Y_i; \tag{2.16}$$

it is “most like” in the sense that it is also a weighted sum of squares where each summand yields an area’s contribution to the total loss that reflects the size of its population. Here, undercount in more populous areas receive more weight, so that using such loss functions reflects an emphasis on national considerations. The loss function $\sum_{i=1}^I (Y_i^{\text{est}} - Y_i)^2 / Y_i^2$, which guarantees undercount equity for the I areas, will not be considered in this article.

It is easy to show that the Bayes estimator in the case of loss function (2.16) is given by,

$$Y_i^{\text{est}} = \left[E \left(\left(\sum_{j=1}^J F_{ji} C_{ji} \right)^{-1} \mid \{X_{ji}: i = 1, \dots, I; j = 1, \dots, J\} \right) \right]^{-1}, \quad (2.17)$$

which is *not* a linear combination of $\{F_{ji}^{\text{uba}}: j = 1, \dots, J\}$. However to a first approximation, using the δ -method, it can be shown that this $Y_i^{\text{est}} \approx Y_i^{\text{uba}}$. This is in fact true for a much larger class of loss functions suggested by Cressie (1987b):

$$L^\lambda \equiv \frac{2}{\lambda(\lambda + 1)} \sum_{i=1}^I \left\{ Y_i^{\text{est}} \left[\left(\frac{Y_i^{\text{est}}}{Y_i} \right)^\lambda - 1 \right] + \lambda [Y_i - Y_i^{\text{est}}] \right\}; \lambda \neq 0, -1; \quad (2.18)$$

the cases $\lambda = 0, -1$ are defined as the respective limits of L^λ as $\lambda \rightarrow 0, -1$. Read and Cressie (1988, Chapter 8) show that in this case the Bayes estimator is

$$Y_i^{\text{est}(\lambda)} = \left[E \left(\left(\sum_{j=1}^J F_{ji} C_{ji} \right)^{-\lambda} \mid \{X_{ji}: i = 1, \dots, I; j = 1, \dots, J\} \right) \right]^{-1/\lambda}, \quad (2.19)$$

which reduces to (2.14) when $\lambda = -1$, and to (2.17) when $\lambda = 1$.

The curious fact is that most undercount estimators used are optimal (under various model assumptions) for $\lambda = -1$, but their performance is measured using $\lambda = 1$; *i.e.*, (2.16). The δ -method argument gives $Y_i^{\text{est}(\lambda)} \approx Y_i^{\text{uba}}$, and recall Y_i^{uba} is optimal for (2.15); therefore squared-error loss estimators of undercount perform well according to a large class of loss functions. This was observed by Kadane (1984) in his hierarchical Bayesian analysis of 1980 census undercount data ($\lambda = -1$ and $\lambda = -2$ were compared), and confirmed on the studies of artificial populations carried out by Cressie and Dajani (1988).

It has just been demonstrated that the estimators (2.13) and (2.14) are Bayes (or approximately so) for a large class of loss functions. However it is not likely that the ensemble properties of $\{F_{ji}^{\text{uba}}: i = 1, \dots, I; j = 1, \dots, J\}$, estimate the corresponding ensemble properties of $\{F_{ji}: i = 1, \dots, I; j = 1, \dots, J\}$, very well. This follows from the inequality $\text{var}(\theta) \geq \text{var}(E(\theta \mid X))$; in other words the posterior mean of the parameter has a smaller variance than the parameter itself. For estimation of state population totals, this does not matter, but for estimation of the *distribution* of say $\{F_{ji} C_{ji}: i = 1, \dots, 51\}; j = 1, \dots, J$, or $\{Y_i: i = 1, \dots, 51\}$, (2.13) is ill-suited to the task. Such a distribution is needed in standards research (Mulry-Liggan and Hogan 1986) to determine the proportion of people in a stratum affected by an undercount more severe than $u\%$ (Cressie 1988, Section 4).

I shall constrain the estimator of $\{F_{ji}: i = 1, \dots, I\}$ so that the posterior moments of its (weighted) empirical distribution function match the moments of the estimator’s weighted empirical distribution function. This is achieved by modifying the usual Bayes estimator, yielding a constrained Bayes estimator with the right ensemble properties. Louis (1984) presents the details for an equal-variance version of the model (2.7), (2.10), but a straightforward modification of his approach is possible for weighted variances. Cressie (1986) shows that such a *constrained Bayes estimator* is

$$F_{ji}^{cba} = \zeta_j + G_j(X_{ji} - \zeta_j), \quad (2.20)$$

$$Y_i^{cba} = \sum_{j=1}^J F_{ji}^{cba} C_{ji}, \quad (2.21)$$

obtained by solving for ζ_j and G_j in:

$$\begin{aligned} \zeta_j + G_j(X_{j\cdot} - \zeta_j) &= F_j + D_j(X_{j\cdot} - F_j); \\ G_j^2 \sum_i \left(C_{ji} / \sum_h C_{jh} \right) (X_{ji} - X_{j\cdot})^2 &= \\ (I-1)D_j\sigma_j^2 / \sum_h C_{jh} + D_j^2 \sum_i \left(C_{ji} / \sum_h C_{jh} \right) (X_{ji} - X_{j\cdot})^2, \end{aligned} \quad (2.22)$$

where

$$X_{j\cdot} = \sum_{i=1}^I X_{ji} C_{ji} / \sum_{h=1}^I C_{jh}. \quad (2.23)$$

2.3 Risks of Adjustment; Model Parameters Assumed Known

The model-based approach described in the previous section specifies undercounts in various area-strata combinations, to be random variables. When it comes to comparing the value of one adjustment procedure against another, the expected loss (or the *risk*) is used. Statistical procedures with small risk are preferred.

In the absence of other considerations (*e.g.*, political, practical, *etc.*), implementing the procedure with the smallest risk is the correct, impartial approach. The statistician knows that adherence to this *modus operandi* will yield better estimates *on the average*, where the average is taken over all problems considered by the statistician. However there is nothing to guarantee that for the particular problem being considered, here estimation of undercount in the 1990 census, a set of area-strata estimates derived from the criterion of minimum risk will actually have smaller loss than another set of estimates. To put it more succinctly, the inequality $E(V^2) < E(W^2)$ does not guarantee that $V^2 < W^2$ for a particular realization. If, in the light of the data collected, a minimum risk prediction did not prove to be the most accurate, the statistical *procedure* should still be seen as optimal.

In the rest of this section, various results about Bayes estimators will be stated (proofs are given in Cressie 1988). Needless to say, these results rely on the correctness of the assumed model. In practice, the more relevant results are for *empirical* Bayes estimators, which are given (with proofs) in Section 3.

The first thing to recall (from Section 2.2) about the usual Bayes estimators (2.13), (2.14) is that they are optimal or near optimal for a large class of loss functions. Moreover the estimators are level-free; *i.e.*, they are not only optimal at the level at which they are constructed, but after aggregation they are also optimal at the higher level. From (2.14),

$$Y_i^{uba} + Y_{i'}^{uba} = Y_{i \& i'}^{uba}, \quad (2.24)$$

where $i \& i'$ denotes the area obtained by combining the two disjoint areas i and i' .

Therefore, one should aim to construct a Bayes estimator at the very lowest level (census blocks) and aggregate up to whatever level is desired, thus ensuring consistency of counts at all levels. In practice this is out of the question, simply because the post-enumeration survey would *never* be large enough to give dual-system estimated undercount data for all the blocks. The same is true at the enumeration-district level and the county level. Moreover, at these lower levels the model (2.7) and (2.10) does not fit as well (Cressie and Dajani 1988); an adequate fit at the state level is shown in Section 3.1.

It is certain that the post-enumeration survey will gather data from each of the 51 states, allowing construction of (empirical) Bayes estimators at the state level. Politically, the state level is the most sensitive; reapportionment of the 50 states' representation (Washington, D.C. is excluded) in the House is the first use made of decennial census counts (mandated to reach Congress by December 31 in the year of the census). Thus at this level, the Bayes estimators (2.13) and (2.14) offer a compromise between a *state's* observed adjustment factors $\{X_{ji}: j = 1, \dots, J\}$; and the (synthetic) adjustment factors $\{F_j: j = 1, \dots, J\}$. For example, Mississippi's black undercount is recognized as being potentially different from New York's black undercount, when using the Bayes estimators.

I shall now explore the consequences of *synthetic* estimation at lower levels, after Bayes estimation is carried out at a given level. For consistency of counts at all levels, it is desired to estimate undercount at the block level and aggregate up to whatever level is desired. Suppose an adjustment factor F_{ji}^{est} is estimated for the j -th stratum in the i -th area. Now suppose $i = i_1 \& i_2$; *i.e.*, the i -th area is split up into two disjoint subareas i_1 and i_2 . Then the synthetic method at the lower level posits,

$$F_{ji_1}^{\text{sy}} = F_{ji_2}^{\text{sy}} = F_{ji}^{\text{est}}, \quad (2.25)$$

so that estimators of the true population are given by,

$$Y_{i_1}^{\text{sy}} = \sum_{j=1}^J F_{ji_1}^{\text{sy}} C_{ji_1}; \quad Y_{i_2}^{\text{sy}} = \sum_{j=1}^J F_{ji_2}^{\text{sy}} C_{ji_2}. \quad (2.26)$$

Notice that from (2.25) and (2.26).

$$Y_{i_1}^{\text{sy}} + Y_{i_2}^{\text{sy}} = Y_i^{\text{est}} \equiv \sum_{j=1}^J F_{ji}^{\text{est}} C_{ji}, \quad (2.27)$$

which is the desired disaggregation-aggregation property.

Compare the risk of using Y_i^{uba} , Y_i^{sya} , and Y_i^{cba} (given by (2.14), (2.5), and (2.21) respectively) to the risk of using C_i , the census count of the i -th area. Using the loss function (2.15), the risks are:

$$\text{uba-risk}_i \equiv E[(Y_i^{\text{uba}} - Y_i)^2 f(C_i)], \quad (2.28)$$

$$\text{cen-risk}_i \equiv E[(C_i - Y_i)^2 f(C_i)], \quad (2.29)$$

$$\text{sya-risk}_i \equiv E[(Y_i^{\text{sya}} - Y_i)^2 f(C_i)], \quad (2.30)$$

$$\text{cba-risk}_i \equiv E[(Y_i^{\text{cba}} - Y_i)^2 f(C_i)]. \quad (2.31)$$

The following sequence of inequalities can be proved (Cressie 1988):

$$\text{uba-risk}_i \leq \text{cba-risk}_i \leq \text{sya-risk}_i \leq \text{cen-risk}_i, \tag{2.32}$$

where the middle inequality requires $\sigma_j^2/\tau_j^2 \leq 3; j = 1, \dots, J$.

Now compare the risk of using $Y_{i_1}^{\text{sy}} \text{ and } Y_{i_2}^{\text{sy}}$ (estimators of Y_{i_1} and Y_{i_2} respectively) based on F_{ji}^{uba} in (2.25), with the risk of using C_{i_1} and C_{i_2} , where area $i = i_1 \text{ \& } i_2$, the union of disjoint areas i_1 and i_2 . It can be shown (Cressie 1988) that the synthetic estimation based on the usual Bayes estimator defined at a particular level but applied at a lower level, always has smaller risk than the census counts.

It is also of interest to determine the behaviour of the census-based risk minus the Bayes-then-synthetic-based risk as a function of the level; the larger this difference, the more advantageous it is to adjust the census counts. Here use $f(C_i) = 1/C_i$ in loss function (2.15). It is possible to show (Cressie 1988) that as disaggregation proceeds to a lower level, the “risk gap” between Bayes-then-synthetic estimation and census counts widens in absolute terms. Although this is proved there for the uba-then-synthetic-based estimator, the same is true for cba-then-synthetic-based and sya-then-synthetic-based estimators, and the ordering of risks (2.32) is preserved at any level of disaggregation. This conclusion depends on the model (2.7) and (2.10) holding at *all* levels. Unfortunately at the lower levels there is some evidence that biases can be substantial. That is, $E(F_{ji}) = F_j + b_{ji}; E(X_{ji} | F_{ji}) = F_{ji} + d_{ji}$. Realistically b_{ji} ’s and d_{ji} ’s are *never* zero, but at sufficiently high levels of aggregation they are unimportant. At the block and enumeration-district level they can be substantial (Cressie and Dajani 1988) and could invalidate the risk inequalities proved so far. Moreover, at lower levels, the data $\{X_{ji}\}$ are more variable leading to less precise estimates of $D_j = \tau_j^2 / (\tau_j^2 + \sigma_j^2)$ in the *empirical* Bayes version (see Section 3) of the Bayes estimator (2.14). These observations, as well as a recognition of the difference between risk and loss, help to explain the deterioration of the performance of the adjusted counts at lower levels, observed in artificial populations (Schultz *et al.* 1986).

3. EMPIRICAL BAYES ADJUSTMENT OF CENSUS COUNTS

Obtain from (2.14), (2.21), and (2.5), the estimated (or adjusted) true area counts Y_i^{uba} , Y_i^{cba} , and Y_i^{sya} , respectively. In order to make these functions only of the data, estimators are needed for the unknown parameters F_j , τ_j^2 , and σ_j^2 ; Fay and Herriot (1979) give empirical Bayes estimators in a regression setting, of which the model (2.7), (2.10) is a special case. For reasons of statistical consistency (see Cressie 1986, Section 3.3), choose,

$$\hat{F}_j = X_j. \tag{3.1}$$

$$\hat{\tau}_j^2 = \max \left\{ \left[\sum_i C_{ji} I(C_{ji} > 0) (X_{ji} - X_{j.})^2 / \left(\sum_i I(C_{ji} > 0) - 1 \right) \right] - \hat{\sigma}_j^2, 0 \right\} \tag{3.2}$$

$\hat{\sigma}_j^2$ is obtained from sampling considerations: it is known for dual-system estimation, and Schultz *et al.* (1986) determine it for their artificial populations by replicating probability-proportional-to-size sampling of 1,440 enumeration districts from the approximately 300,000 total number.

Statistical stability (*i.e.*, small sampling variance) for sample means is easier to achieve than for sample variances. The coefficient of variation of the sample variance is approximately $\sqrt{2}/\sqrt{n}$; therefore to achieve a relative confidence region (0.5, 1.5) for the population variance, a value of $n = 32$ is needed; and to achieve a region (0.95, 1.05) a value of $n = 3,200$ is needed. Thus the estimator, $\sum_{i=1}^I C_{ji} I(C_{ji} > 0) (X_{ji} - X_{j\cdot})^2 / (\sum_{i=1}^I I(C_{ji} > 0) - 1)$ of $\tau_j^2 + \sigma_j^2$ is very unstable, particularly when there are a large number of strata and hence $\sum_{i=1}^I I(C_{ji} > 0)$ is small (smaller than 30).

One way around this is to introduce a further mixing distribution into the problem, namely, model the $\{\tau_j^2: j = 1, \dots, J\}$ as being generated by the reciprocal of a gamma distribution for example. Thus instead of estimating J parameters $\{\tau_j^2: j = 1, \dots, J\}$, the problem can be reduced to estimating just two gamma parameters (see *e.g.*, Hui and Berger 1983). Another possibility is to aggregate temporarily some of the strata for the purpose of estimating the stratum variance. In other words, define disjoint groups of strata indices, A_1, \dots, A_K , such that $\cup \{A_k: k = 1, \dots, K\} = \{1, 2, \dots, J\}$, and $\tau_j^2 = \tau_{j'}^2 = T_k^2$, whenever j and j' belong to the same A_k . In this way, Cressie and Dajani (1988) reduce the number of stratum variance parameters from $J = 96$ down to $K = 4$. For the data analyzed below, since $\sum_{i=1}^I I(C_{ji} > 0) = 51$ for each of the three race strata, it was not necessary to "borrow strength" in the ways just described.

3.1 Empirical Bayes Estimators

The usual (see, for example, Morris 1983) and constrained (Louis 1984) empirical Bayes estimators can now be constructed:

$$F_{ji}^{\text{ueb}} = X_{j\cdot} + \{\hat{\tau}_j^2 / (\hat{\tau}_j^2 + \hat{\sigma}_j^2)\} (X_{ji} - X_{j\cdot}), \quad (3.3)$$

$$Y_i^{\text{ueb}} = \sum_{j=1}^J F_{ji}^{\text{ueb}} C_{ji}; i = 1, \dots, I; \quad (3.4)$$

$$F_{ji}^{\text{ceb}} = X_{j\cdot} + \{\hat{\tau}_j^2 / (\hat{\tau}_j^2 + \hat{\sigma}_j^2)\}^{1/2} (X_{ji} - X_{j\cdot}), \quad (3.5)$$

$$Y_i^{\text{ceb}} = \sum_{j=1}^J F_{ji}^{\text{ceb}} C_{ji}; i = 1, \dots, I. \quad (3.6)$$

The usual empirical Bayes estimator (3.3) can also be obtained from standard theory for linear models with random effects (Henderson 1976).

Notice that when $\hat{\tau}_j^2 = 0$, the empirical Bayes estimators of the j -th stratum adjustment factors all reduce to the synthetic estimator $X_{j\cdot}$. The presence of the weight $\{\hat{\tau}_j^2 / (\hat{\tau}_j^2 + \hat{\sigma}_j^2)\}^{1/2}$ in the constrained empirical Bayes estimator (3.5) may look a little strange at first, but it is seen in Cressie (1987a) to yield an unbiased estimator of the stratum error $C_{ji}^{1/2} (F_{ji} - F_j)$.

An earlier suggestion for empirical Bayes modeling of undercount came from Dempster and Tomberlin (1980), who proposed that the number of undercounted people in a subarea might be a binomial random variable. They defined a heirarchical Bayes model but did not take into account the heteroskedastic variation. Stroud (1987) introduces a covariate into a two-stage Bayesian model, but his assumptions of homoskedastic variation and equal sample sizes in each subarea, are too restrictive for the problem considered in this article.

Formulas for the bias and mean-squared error of the usual empirical Bayes (ueb) estimators (3.3), (3.4), the constrained empirical Bayes (ceb) estimators (3.5), (3.6), and the synthetic estimators

$$F_{ji}^{syn} = X_j. \tag{3.7}$$

$$Y_i^{syn} = \sum_{j=1}^J F_{ji}^{syn} C_{ji}; i = 1, \dots, I, \tag{3.8}$$

are given in Cressie (1987a, Section 4). Since undercount is a nonlinear function of the true population, its estimators based on $\{F_{ji}^{est}; i = 1, \dots, I; j = 1, \dots, J\}$, viz.

$$u_{ji}^{est} \equiv 1 - \frac{1}{F_{ji}^{est}}; i = 1, \dots, I; j = 1, \dots, J, \tag{3.9}$$

$$u_i^{est} \equiv 1 - \frac{C_i}{Y_i^{est}}; i = 1, \dots, I, \tag{3.10}$$

are biased; estimated biases and mean-squared errors can be obtained by the δ -method (Cressie 1987a, Section 4). All of these bias and mean-squared error calculations do not take into account variation due to the (nonlinear) estimation of $\tau_j^2 / (\tau_j^2 + \sigma_j^2)$.

Suppose that the following three U.S. strata (based on race/ethnicity) are chosen: blacks, nonblack hispanics, and others. Data from the post-enumeration survey following the 1980 U.S. Census are given in Cressie (1987a, Table 1). These are from the *noninstitutional* population (Cowan and Bettin 1982) and have been labeled “PEP 3-8” by the U.S. Census Bureau – the “3” refers to census omissions being obtained from an April survey and to imputing missing data, and the “8” refers to erroneous enumerations being obtained from a separate survey that imputed missing data with the help of U.S. Post Office information.

From these data and (3.1), (3.2), Cressie (1987a) estimated the mean of the mixture distribution, and standardized stratum and sampling variances defined in (2.7) and (2.10):

blacks: $\hat{F}_1 = 1.06076 \quad \hat{\tau}_1^2 = 673.982 \quad \hat{\sigma}_1^2 = 522.183,$ (3.11)

nonblack
hispanics: $\hat{F}_2 = 1.04667 \quad \hat{\tau}_2^2 = 308.990 \quad \hat{\sigma}_2^2 = 246.585,$ (3.12)

Others: $\hat{F}_3 = 0.99981 \quad \hat{\tau}_3^2 = 242.134 \quad \hat{\sigma}_3^2 = 242.152.$ (3.13)

Based on these parameter estimators and the PEP 3-8 data $\{X_{ji}; j = 1, 2, 3; i = 1, \dots, 51\}$, Cressie (1987a) gave undercount estimates $\{u_{ji}^{est}\}$, $\{u_i^{est}\}$ for ueb-based and syn-based estimators defined by (3.3) and (3.7) respectively.

To check the fit of the model, the residuals $\{C_{ji}^{1/2} (F_{ji}^{ceb} - F_j^{ceb}); i = 1, \dots, I\}$ were computed for each of the three strata. Table 1 shows the results, presented as stem-and-leaf plots for the three race strata; a bell-shaped plot for each is the ideal. The model appears to fit the data, except for the nonblack-hispanic stratum in the state of New York. In light of the lawsuit, Cuomo vs. Baldrige, heard by the Southern District Court of New York in 1983, this new way of looking at the data tells an interesting story. The nonblack hispanics in New York State

Table 1
Stem and leaf plots of residuals based on "ceb" estimator

Blacks ($j = 1$)			Nonblack ($j = 2$)			Others ($j = 1$)		
STEM	LEAF	#	STEM	LEAF	#	STEM	LEAF	#
5	2	1	7	1	1	4	2	1
4	58	2	6			3	8	1
3	048	3	5			3		
2	8	1	4			2		
1	259	3	3	1	1	2	012	3
0	114566788	9	2	15	2	1	55568	5
-0	87665444322100	15	1			1	22	2
-1	83310	5	0	111235599	9	0	55556889	8
-2	75310	5	-0	99987755532221111100	21	0	1112344	7
-3	54	2	-1	55410000	8	0	3200	4
-4	7	1	-2	4333322	7	-0	98777655	8
-5	9872	4	-3	4	1	-1	31000	5
MULTIPLY STEM.LEAF BY 10**+01			-4	1	1	-1	5	1
			MULTIPLY STEM.LEAF BY 10**+01			-2	211	3
						-2	66	2
						-3		
						-3		
						-4	2	1
						MULTIPLY STEM.LEAF BY 10**+01		

were grossly undercounted, even in relation to their undercounted fellow nonblack hispanics in other states. Incidentally, the judge decided in favour of the U.S. Department of Commerce (in December 1987) on the grounds that the statistical and demographic professions had not developed adequate methods of adjustment for the whole country by 1980.

When are census counts improved by replacing $\{C_i; i = 1, \dots, I\}$ with $\{Y_i^{\text{est}}; i = 1, \dots, I\}$? The next section gives conditions under which an analogous ordering to (2.32) still holds in the *empirical* Bayes setting.

3.2 Adjustment at Different Levels; Model Parameters Estimated

The same comments at the beginning of Section 2.3 apply; in a model-based approach a small risk does not guarantee a small loss in every problem but only on the average. Also the analogous aggregation property to (2.24) holds for ueb-based, ceb-based, and syn-based estimators, namely

$$Y_i^{\text{est}} + Y_{i'}^{\text{est}} = Y_{i \& i'}^{\text{est}}, \tag{3.14}$$

for “est” = “ueb,” “ceb,” and “syn,” given by (3.4), (3.6), and (3.8) respectively. Moreover the disaggregation-aggregation property (2.27), namely

$$Y_{i_1}^{\text{syce}} + Y_{i_2}^{\text{syce}} = Y_i^{\text{est}}, \tag{3.15}$$

where $i = i_1 \& i_2$ and $F_{ji_1}^{\text{syce}} = F_{ji_2}^{\text{syce}} = F_{ji}^{\text{est}}$, holds for any estimator of F_{ji} , including those based on ueb, ceb, and syn.

Write the risk of estimating Y_i by Y_i^{est} ($= \sum_{j=1}^J F_{ji}^{\text{est}} C_{ji}$) as

$$\text{est-risk}_i \equiv E[(Y_i^{\text{est}} - Y_i)^2 f(C_i)]. \tag{3.16}$$

The estimators given by “est” = “ueb,” “ceb,” and “syn,” will be compared to “cen” ($F_{ji}^{\text{cen}} \equiv 1$) via (3.16). For the rest of this section consider the estimator,

$$F_{ji}^{\text{est}} = r_j X_{ji} + (1 - r_j) X_{j.}; \quad 0 \leq r_j \leq 1, \tag{3.17}$$

a convex combination of the data X_{ji} and the synthetic estimator $X_{j.}$. Then

$$\text{est-risk}_i = \sum_{j=1}^J \tau_j^2 (1 - r_j)^2 \left\{ C_{ji} - \frac{C_{ji}^2}{\sum_h C_{jh}} \right\} + \sigma_j^2 \left\{ r_j^2 C_{ji} + \frac{(1 - r_j^2) C_{ji}^2}{\sum_h C_{jh}} \right\}. \tag{3.18}$$

It is easy to see that the value of r_j that minimizes (3.18) is $r_j = D_j = \tau_j^2 / (\tau_j^2 + \sigma_j^2)$; *i.e.*, neglecting the effect of estimating τ_j^2 and σ_j^2 , I obtain

$$\text{ueb-risk}_i \leq \text{est-risk}_i; \quad 0 \leq r_j \leq 1. \tag{3.19}$$

Now compare ueb-risk_{*i*} (put $r_j = D_j$ in (3.17)) with cen-risk_{*i*}; recall from (2.29)

$$\text{cen-risk}_i = \sum_{j=1}^J \tau_j^2 C_{ji} f(C_i) + \left[\sum_{j=1}^J (F_j - 1) C_{ji} \right]^2 f(C_i). \quad (3.20)$$

Also, by putting $\tau_j^2 = k_j \sigma_j^2$; $j = 1, \dots, J$,

$$\text{ueb-risk}_i = \sum_{j=1}^J \sigma_j^2 \left\{ \frac{k_j}{1 + k_j} + \frac{C_{ji}}{\sum_h C_{jh}} \cdot \frac{1}{1 + k_j} \right\} C_{ji} f(C_i). \quad (3.21)$$

A sufficient condition for ueb-risk_{*i*} ≤ cen-risk_{*i*} is,

$$\left\{ \frac{k_j}{1 + k_j} + \frac{C_{ji}}{\sum_h C_{jh}} \cdot \frac{1}{1 + k_j} \right\} \leq k_j;$$

that is, if

$$\sigma_j^2 / \tau_j^2 \leq \left\{ \sum_h C_{jh} / C_{ji} \right\}^{1/2}; j = 1, \dots, J, \quad (3.22)$$

then

$$\text{ueb-risk}_i \leq \text{cen-risk}_i. \quad (3.23)$$

Similarly, it can be shown that if

$$\sigma_j^2 / \tau_j^2 \leq 1; j = 1, \dots, J, \quad (3.24)$$

then

$$\text{syn-risk}_i \leq \text{cen-risk}_i. \quad (3.25)$$

Finally, if $(\sigma_j^2 / \tau_j^2) \leq 1$, and

$$4(\sigma_j^2 / \tau_j^2)^2 \left(\frac{C_{ji}}{\sum_h C_{jh}} \right)^2 - (\sigma_j^2 / \tau_j^2) \left(1 + \frac{2C_{ji}}{\sum_h C_{jh}} \right) + 3 \geq 0; j = 1, \dots, J, \quad (3.26)$$

then

$$\text{ceb-risk}_i \leq \text{syn-risk}_i. \quad (3.27)$$

once again (from (3.26)), if σ_j^2 / τ_j^2 is small, risks can be bounded.

Therefore an analogous sequence of inequalities to (2.32) is possible:

$$\text{ueb-risk}_i \leq \text{ceb-risk}_i \leq \text{syn-risk}_i \leq \text{cen-risk}_i, \tag{3.28}$$

where the middle inequality requires the condition (3.26) and the last inequality requires the condition (3.24). If either of these two inequalities do not hold, at least the ueb-based estimator is an improvement over the census counts if condition (3.22) is satisfied. For the PEP 3-8 data from the 1980 U.S. Census,

$$\hat{\sigma}_1^2/\hat{\tau}_1^2 = 0.77, \hat{\sigma}_2^2/\hat{\tau}_2^2 = 0.80, \hat{\sigma}_3^2/\hat{\tau}_3^2 = 1.00; \tag{3.29}$$

that is, for the 1980 U.S. decennial census the census risk is larger than the synthetic risk and the usual-empirical-Bayes risk is smallest of all.

Now compare the risk of using $Y_{i_1}^{\text{syce}}$ and $Y_{i_2}^{\text{syce}}$ (estimators of Y_{i_1} and Y_{i_2} respectively, based on F_{ji}^{est} given by (3.17)), with the risk of using C_{i_1} and C_{i_2} , where area $i = i_1$ & i_2 is disaggregated into two disjoint areas i_1 and i_2 .

$$\begin{aligned} & \sum_{\ell=1}^2 E \left[(Y_{i_\ell}^{\text{syce}} - Y_{i_\ell})^2 f(C_{i_\ell}) \right] \\ &= \sum_{\ell=1}^2 \sum_{j=1}^J \left[\tau_j^2 \left\{ (1 - r_j)^2 \left(\frac{1}{C_{ji}} - \frac{1}{\sum_h C_{jh}} \right) + \left(\frac{1}{C_{ji_\ell}} - \frac{1}{C_{ji}} \right) \right\} \right. \\ & \quad \left. + \sigma_j^2 \left\{ \frac{1 - r_j^2}{\sum_h C_{jh}} + \frac{r_j^2}{C_{ji}} \right\} \right] C_{ji_\ell}^2 f(C_{i_\ell}). \end{aligned} \tag{3.30}$$

It is easy to see that under precisely the same conditions (3.22), (3.24), (3.26), the same sequence of inequalities (3.28) holds; interpret est-risk_i in (3.28) as being equal to (3.30) with $r_j = D_j$ for “est” = “ueb,” with $r_j = D_j^{1/2}$ for “est” = “ceb,” and with $r_j = 0$ for “est” = “syn”. Moreover for the loss function (2.15) with $f(C_i) = 1/C_i$, risk gaps widen as lower levels of aggregation are attained.

4. DISCUSSION

Various assumptions are made in deriving the risk inequalities (3.28), all of which deserve further investigation. The model (2.7) and (2.10) is assumed to fit, and in particular the independence of distributions between subareas is assumed. Moreover, the effect of estimating D_j in the empirical Bayes estimators of F_{ji} is assumed negligible. Notice however that synthetically estimated F_{ji} ’s do not use an estimate of D_j and so those risk inequalities only rely on the appropriateness of the model (2.7), (2.10).

The conditions which order the various risks and bound them below the census risk in (3.28), all depend on σ_j^2/τ_j^2 being “small.” The practical implication is that a large number of households need to be chosen in the post-enumeration survey (PES) or there can be no guarantee that census counts can be improved by adjustment. With prior knowledge of stratum variation (e.g., from a previous census), the PES could be *designed* so that the conditions are satisfied.

After the survey has been conducted and the data $\{X_{ji}; i = 1, \dots, I; j = 1, \dots, J\}$ are available, the various conditions (3.22), (3.24), and (3.26) can all be checked by using the estimators $\hat{\tau}_j^2$ and $\hat{\sigma}_j^2$ given by (3.2).

Concentrate on the best convex combination of X_{ji} and $X_{j\cdot}$, namely F_{ji}^{ueb} given by (3.3). Then, $\text{ueb-risk}_i \leq \text{cen-risk}_i$, if (3.22) holds; i.e., if

$$\sigma_j^2 / \tau_j^2 \leq \left\{ \sum_h C_{jh} / C_{ji} \right\}^{1/2}; j = 1, \dots, J. \quad (4.1)$$

Notice that the condition is less stringent when the i -th area has a small census population; conversely, areas of large census population may have a ueb-based estimated population further from the truth than census. A sufficient condition for (4.1) to hold is, $\sigma_j^2 / \tau_j^2 \leq 1$; $j = 1, \dots, J$, which is also the condition that guarantees the syn-based estimated population improves over census. This condition was satisfied for the 1980 PEP 3-8 data (see Section 3.2).

Finally, the condition (4.1) becomes less stringent at lower levels, and indeed the results of Section 3.2 show that the risk gap between the adjusted population and the census population widens. This deserves comment. The results are true provided the model holds at lower levels, but this is probably not the case at the block and the enumeration-district level. Presence of bias in (2.7) and (2.10); namely

$$E(F_{ji}) = F_j + b_{ji}; E(X_{ji} | F_{ji}) = F_{ji} + d_{ji}, \quad (4.2)$$

could cause a reversal in some of the risk inequalities. At the state level however, Table 1 and Cressie (1988) show through an examination of residuals, that (2.7) and (2.10) does fit for the 1980 PEP 3-8 data. And since (3.29) implies that condition (4.1) is satisfied, one can be confident that ueb-based adjusted state totals are closer to the truth than census state totals. That may not be true at the block level; clearly a decision regarding the level at which it is most important to have accurate census counts, needs to be made. The first use of U.S. Census data is the reporting of *state* totals to Congress for the purpose of redistricting House seats. One might include a number of large cities in with the states, and create e.g., the "states" New York City, and New York State Except New York City. It seems to me that this "state" level is the most sensitive politically and that accurate totals at this level should receive the highest priority.

ACKNOWLEDGEMENT

This research has benefited from the input of members of the Undercount Research Staff at the U.S. Bureau of the Census, and from the perceptive comments of J.B. Kadane and J.W. Tukey. Support from Joint Statistical Agreements JSA 86-5, JSA 87-4, JSA 87-10, and JSA 88-13, between Iowa State University and the Census Bureau is gratefully acknowledged. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau.

REFERENCES

- COWAN, C.D., and BETTIN, P.J. (1982). Estimates and missing data problems in the post enumeration survey. Internal Report. Statistical Methods Division, Bureau of the Census, Washington, D.C.
- CRESSIE, N. (1986). Empirical Bayes estimation of undercount in the decennial census. *Statistical Laboratory Preprint 86-58*, Iowa State University, Ames, IA.
- CRESSIE, N. (1987a). Empirical Bayes estimation of undercount in the decennial census. Manuscript submitted to *Journal of the American Statistical Association*.
- CRESSIE, N. (1987b). Comment on "Census undercount adjustment and the quality of geographic population distributions," by A.L. Schirm and S.H. Preston. *Journal of the American Statistical Association*, 82, 980-983.
- CRESSIE, N. (1988). Estimating census undercount at national and subnational levels. *Proceedings of Bureau of the Census Fourth Annual Research Conference*. Bureau of the Census, Washington, D.C., 123-150.
- CRESSIE, N., and DAJANI, A. (1988). Empirical Bayes estimation of U.S. undercount based on artificial populations. *Statistical Laboratory Preprint 88-17*, Iowa State University, Ames, IA.
- DEMPSTER, A.P., and TOMBERLIN, T.J. (1980). The analysis of census undercount from a post-enumeration survey, in *Proceedings of the 1980 Conference on Census Undercount*. Bureau of the Census, Washington, D.C. 88-94.
- ERICKSEN, E.P., and KADANE, J.B. (1985). Estimating the population in a census year: 1980 and beyond. *Journal of the American Statistical Association*, 80, 98-131.
- ERICKSEN, E.P., and KADANE, J.B. (1987). Sensitivity analysis of local estimates of undercount in the 1980 U.S. Census, in *Small Area Statistics*, (Eds. R. Platek, J.N.K. Rao, C.E. Särndal and M.P. Singh.) New York: Wiley, 23-45.
- ERICKSEN, E.P., KADANE, J.B., and TUKEY, J.W. (1987). Adjusting the 1980 census of housing and population. *Technical Report No. 401*, Department of Statistics, Carnegie-Mellon University, Pittsburgh, PA.
- FAY, R.E. III, and HERRIOT, R.A. (1979). Estimates of income for small places: An application of James-Stein to census data. *Journal of the American Statistical Association*, 74, 269-277.
- FERGUSON, T.S. (1967). *Mathematical Statistics: A Decision Theoretic Approach*. New York: Academic Press.
- FREEDMAN, D.A., and NAVIDI, W.C. (1986). Regression models for adjusting the 1980 census. *Statistical Science*, 1, 3-39.
- GOLDSTEIN, M. (1975). Approximate Bayes solutions to some nonparametric problems. *Annals of Statistics*, 3, 512-517.
- HENDERSON, C.R. (1976). A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics*, 32, 69-83.
- HUI, S.L., and BERGER, J.O. (1983). Empirical Bayes estimation of rates in longitudinal studies. *Journal of the American Statistical Association*, 78, 753-760.
- ISAKI, C.T., DIFFENDAL, G.J., and SCHULTZ, L.K. (1986). Statistical synthetic estimates of undercount for small areas. *Proceedings of Bureau of the Census Second Annual Research Conference*. Bureau of the Census, Washington, D.C., 557-569.
- KADANE, J.B. (1984). Allocating Congressional seats among the states when state populations are uncertain. *Technical Report No. 309*, Department of Statistics, Carnegie-Mellon University, Pittsburgh, PA.
- LINDLEY, D.V., and SMITH, A.F.M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B*, 34, 1-41.
- LOUIS, T.A. (1984). Estimating a population of parameter values using Bayes and empirical Bayes methods. *Journal of the American Statistical Association*, 79, 393-398.

- MORRIS, C.N. (1983). Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association*, 78, 47-55.
- MULRY-LIGGAN, M., and HOGAN, H. (1986). Research plan on census adjustment standards. *Proceedings of Bureau of the Census Second Annual Research Conference*. Bureau of the Census, Washington, D.C., 381-392.
- NATIONAL ACADEMY OF SCIENCES (1985). *The Bicentennial Census: New Directions for Methodology in 1990*, (Eds. C.F. Citro and M.L. Cohen.) Washington: National Academy Press.
- READ, T.R.C., and CRESSIE, N.A.C. (1988). *Goodness-of-fit Statistics for Discrete Multivariate Data*. New York: Springer-Verlag.
- SCHULTZ, L.K., HUANG, E.T., DIFFENDAL, G.J., and ISAKI, C.T. (1986). Some effects of statistical synthetic estimation on census undercount of small areas. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 321-325.
- STROUD, T.W.F. (1987). Bayes and empirical Bayes approaches to small area estimation, in *Small Area Statistics*, (Eds. R. Platek, J.N.K. Rao, C.E. Särndal and M.P. Singh.) New York: Wiley, 124-137.
- TUKEY, J.W. (1981). Discussion of "Issues in adjusting for the 1980 census undercount," by Barbara Bailar and Nathan Keyfitz, presented at the Annual Meeting of the American Statistical Association, Detroit, MI.

Imputation Strategies for Missing Values in Post-Enumeration Surveys

DONALD B. RUBIN, JOSEPH L. SCHAFER, AND NATHANIEL SCHENKER¹

ABSTRACT

To estimate census undercount, a post-enumeration survey (PES) is taken, and an attempt is made to find a matching census record for each individual in the PES; the rate of successful matching provides an estimate of census coverage. Undercount estimation is performed within poststrata defined by geographic, demographic, and housing characteristics, X . Portions of X are missing for some individuals due to survey nonresponse; moreover, a match status Y cannot be determined for all individuals. A procedure is needed for imputing the missing values of X and Y . This paper reviews the imputation methods used in the 1986 Test of Adjustment Related Operations (Schenker 1988) and proposes two alternative model-based methods: (1) a maximum-likelihood contingency-table estimation procedure that ignores the missing-data mechanism; and (2) a new Bayesian contingency table estimation procedure that does not ignore the missing-data mechanism. The first method is computationally simpler, but the second is preferred on conceptual and scientific grounds.

KEY WORDS: Bayesian methods; Categorical data; Coverage error; EM algorithm; Multiple imputation; Nonignorable nonresponse; Undercount.

1. INTRODUCTION

The U.S. Bureau of the Census has used a post-enumeration survey (PES) to evaluate coverage error in several past censuses, and it plans to conduct a PES after the 1990 Decennial Census as well. For each individual in the PES, an attempt is made to find a census record (*i.e.*, a match) to determine whether the person was enumerated in the census. The proportion of PES persons who were missed in the census is used as an estimate of the proportion of persons in the population who were missed. A similar matching operation is performed to match a sample of individuals from the census to the PES; this provides an estimate of the census overcount resulting from erroneous (*e.g.*, duplicate or fictitious) enumerations.

The data on matches and erroneous enumerations obtained from the PES are combined to estimate the population size via the dual-system estimator; this capture-recapture type of estimator is discussed in Marks, Seltzer and Krotki (1974), Krotki (1978), Wolter (1986), Diefendal (1988), and Fay, Passell and Robinson (1988, Chapter 5). Dual-system estimates of population size are computed within poststrata defined by geographic, demographic (age, sex, race), and housing (owner/renter, type of housing structure) characteristics.

Two problems of missing data occur in the PES and complicate the estimation process:

1. Geographic, demographic, or housing characteristics may be missing for a person, so it is not known to which poststratum that person belongs.
2. After the processing of the PES, there are some individuals with match status (dichotomous variable indicating matched/not matched to census) or erroneous enumeration status missing. This can occur, for instance, when an incomplete name is obtained in the PES, or when there is difficulty in specifying a Census Day address for someone who moved between Census Day and the PES.

¹ Donald B. Rubin and Joseph L. Schafer, Department of Statistics, Harvard University, Cambridge, MA 02138, USA; Nathaniel Schenker, Division of Biostatistics, UCLA School of Public Health, Los Angeles, CA 90024, USA.

Missing data were a major source of uncertainty in undercount estimation for the 1980 Decennial Census (Freedman and Navidi 1986; Fay, Passell and Robinson 1988, Chapter 6). Improvements in the PES design should reduce the amount of missing data in 1990 (Hogan and Wolter 1988), but a method for dealing with missing data will still be necessary.

The 1986 Test of Adjustment Related Operations (TARO), a recent test of undercount estimation and adjustment (Diffendal 1988; Schenker 1988), used a PES that was similar in design to that planned for 1990. This paper reviews the methods used to handle missing data in TARO (Schenker 1988), identifies potential weaknesses of these methods, and discusses potential alternatives.

Our goal is to indicate issues and problems, and to suggest methods for their solution. The long range plan for research is to carefully evaluate these methods. Although we only discuss imputation for missing PES data when estimating undercount, missing data also occur in the census sample used to estimate overcount. The missing-data problems in estimating overcount, however, are analogous to those in estimating undercount (Schenker 1988), and so our discussion applies to both problems.

In our discussion of alternatives to the TARO procedures, we propose a new method based on a Bayesian model that does not ignore the missing-data mechanism, and thus does not assume that the missing data are missing at random. Nonignorable models for incomplete categorical data are a recent development in the theory of handling missing data; see Fay (1986), Little and Rubin (1987, Section 11.6), and Baker and Laird (1988) for discussions and reviews of the literature. Moreover, the types of missing data that we discuss occur not only in undercount estimation, but in many other situations as well; thus our discussion is relevant to the general problem of handling missing categorical data.

Section 2 discusses the imputation methods used in TARO. In Section 3, alternative methods are described and illustrated using a simple example. Section 4 presents a concluding discussion.

2. IMPUTATION METHODS USED IN TARO

2.1 Description of Methods

For each individual in the PES, let X denote categorical variables for age, sex, race, owner/renter status, and type of housing structure; let Y denote match status (1 = match, 0 = nonmatch); and let Z denote variables indicating whether the PES interview was with a household member or a proxy, and whether the PES person moved between Census Day and the PES. In TARO, the X variables (except type of housing structure) were used in forming poststrata (Diffendal 1988); Z was observed for all PES individuals, but Y and components of X were sometimes missing (Schenker 1988).

Missing values of X and Y were imputed in two stages. (Our description is simplified for ease of presentation; see Schenker (1988) for the precise procedure). First, all missing X values were imputed using a "hot deck" scheme based on observed X variables; that is, imputed values were drawn from the observed distributions of X values. Second, after the missing values of X were filled in, a logistic regression model predicting Y from X and Z was fitted to the cases with Y observed. This logistic regression model was then used to impute probabilities of match for all missing Y values. Probabilities rather than zeros and ones were imputed to (a) increase the precision of estimation, and (b) allow the assessment of variability due to imputation (Schenker 1989).

2.2 Critique of Methods

The TARO imputation methods have many positive features. They are easily understood and use explicit modeling for the imputation of Y . They also condition on much of the observed data, rather than imputing from marginal distributions. Finally, in principle they allow the assessment of uncertainty in undercount estimates due to the missing Y values. The methods have some potential weaknesses, however, which we now describe.

The TARO imputation procedure is an "ignorable" procedure, because it ignores the missing-data mechanism. Ignorable procedures assume that the missing data are missing at random (MAR) (Rubin 1976); that is, they assume that given the observed data, the missingness is independent of the values of the missing items. For example, if X and Z are observed for all people, MAR implies that Y can be imputed using the conditional distribution of Y given X and Z for those individuals having X , Y , and Z observed.

The TARO procedure is actually a special case of an ignorable procedure, because it makes assumptions that are stronger than the general MAR assumption. The TARO procedure treated X and Y asymmetrically; that is, it imputed missing values of Y conditional on all observed data, but it imputed missing X values conditional only on the observed X 's, rather than on the observed values of X , Y , and Z . Hence, in addition to the general MAR assumption, the TARO procedure also effectively assumed that, given the observed components of X , the missing components of X are conditionally independent of both Y and Z .

This additional independence assumption may not be realistic; it may be that given the observed X data, there is a residual dependence of values of missing components of X on Y and/or Z . If this is the case, then observed values of Y and Z should be used in the imputation of X . For instance, suppose a PES individual has sex missing, but is found not to match any census record ($Y = 0$) on the basis of observed age, race, and address; and suppose males tend to be undercounted in the census more than females with identical other characteristics. Then knowing that $Y = 0$ provides some evidence that the person in question is more likely to be male than if Y were 1. The most general ignorable imputation procedure would use information provided by Y and Z in imputing missing X values; this is one of the alternative imputation methods, which we discuss in Section 3.4.1.

Another feature of the TARO procedure that may be unrealistic is the ignorability assumption itself. It may be that the missing data are not MAR — *i.e.*, given the observed data, the missingness is not independent of the values of the missing items; if so, then it would be more appropriate to use a nonignorable model for the missing-data mechanism. For instance, consider a group of people with identical values of all variables except race; it may be more difficult to obtain information on race for minorities than nonminorities, and consequently the distribution of race will be different among those missing race and those with race observed. Similarly, even after all X and Z variables are controlled for, it may be that people who were not enumerated in the census are more likely to be missing Y than those who were enumerated in the census. An alternative imputation method based on a general class of nonignorable models is presented in Section 3.4.2.

3. ALTERNATIVE METHODS OF IMPUTATION IN THE PES

3.1 Introduction

Let $X = (X_1, X_2, X_3)$ denote three individual characteristics recorded by the PES (*e.g.*, age, sex, and race). The variables X_1 , X_2 , and X_3 are assumed to be categorical, taking I , J , and K possible values respectively. We have chosen three variables merely for illustrative purposes

and notational simplicity; all ideas developed here will extend immediately to any number of categorical variables. In practice, these X variables will probably include the demographic, geographic, and housing characteristics used to define poststrata for undercount estimation; they may also include additional PES variables, such as mover status and household member/proxy status, which are not of intrinsic interest but which may be useful for imputation purposes.

We will form IKJ different classes of individuals by cross-classifying them according to X_1 , X_2 , and X_3 . These classes may or may not be the same as the poststrata for undercount estimation; in practice the poststrata will probably be coarser than these classes. It is convenient, but not necessary, for these classes to be defined as cross-classifications of all possible values of X_1 , X_2 , and X_3 ; more complicated patterns (such as nested ones) are also possible. We will be constructing loglinear models for cross-classified contingency tables, but loglinear models may be based on other patterns as well.

Let Y be the dichotomous variable denoting match status, taking values 1 (matched to census) or 0 (not matched). If there were no missing data, the results of the PES could be summarized in a single four-dimensional contingency table with $I \times J \times K \times 2$ cells, since each individual could be fully classified according to X_1 , X_2 , X_3 , and Y . But those individuals missing one or more variables can be only partially classified according to those variables that are observed. Those having X_1 , X_2 , X_3 , and Y all observed will constitute a four-dimensional table, which we will call the table of *complete cases* (CC), or the data table for missingness pattern 1 (no variables missing). Those having X_1 , X_2 , and X_3 observed but Y missing will constitute a three-dimensional *supplementary table* with IKJ cells, which we will call the data table for missingness pattern 2. In general, there will be 2^4 such tables corresponding to all possible missingness patterns, one CC table and $2^4 - 1$ supplementary tables.

3.2 Imputation from Reference Tables

In our model-based approach to imputation, we will model the data tables for different missingness patterns as multinomial observations. Corresponding to each missingness pattern, we will define a set of cell probabilities $\Theta^t = \{\Theta^t_{ijkl}\}$, where the superscript t indexes the missingness pattern, $t = 1, \dots, 2^4$, and the subscripts i, j, k , and l indicate the levels of X_1 , X_2 , X_3 , and Y respectively. Because we will refer to Θ^t when imputing missing values for the t -th data table, we will call Θ^t the reference table for the t -th data table, and $\{\Theta^t: t = 1, \dots, 2^4\}$ the set of reference tables.

Imputation of missing values corresponds to expanding each supplementary data table to make it fully four-dimensional, according to its corresponding reference table. For example, consider the imputation of Y for those individuals missing only Y . This is equivalent to expanding the supplementary data table for missingness pattern 2, by dividing each cell count in this table into two parts, a count of those having $Y = 1$ and a count of those having $Y = 0$, split according to the reference table Θ^2 . With known Θ^2 this procedure is straightforward: we first obtain from Θ^2 the conditional distribution of Y given X for this missingness pattern, *i.e.*,

$$P(Y = 1 \mid X_1, X_2, X_3, t = 2) = \frac{\theta^2_{ijk1}}{\theta^2_{ijk0} + \theta^2_{ijk1}}, \tag{1}$$

for $i = 1, \dots, I, j = 1, \dots, J$, and $k = 1, \dots, K$. Then, we impute $Y = 1$ for each observation in cell ijk of this table with probability given by the right-hand side of (1); alternatively, we could impute the mean of this distribution, which is just the probability of a match (1). The relative merits of random draw versus mean imputation for the PES will be discussed in Section 3.3.

Note that in the example above, the only information from Θ^2 needed for the imputation is the conditional distribution of Y given X ; hence, any value of Θ^2 yielding the same values for (1) leads to the same imputation procedure. For an imputation procedure to be accurate, then, our estimate of Θ' need not correspond to the joint distribution of Y and X for the t -th missingness pattern; the only requirement is that the conditional distribution of the missing variables given the observed ones derived from our estimate of Θ' be close to the correct one.

In particular, if the missing-data mechanism is ignorable, one common reference table $\Theta' = \Theta$, $t = 1, \dots, 2^4$, provides valid imputations for all missingness patterns, even though the joint distribution of X and Y might vary across missingness patterns. The fact that only one reference table is needed follows from the definition of ignorability, which implies that the conditional distribution of missing values given observed values does not depend on the missingness pattern. The value Θ that provides valid imputations is not Θ_{CC} , the cell probabilities for the joint distribution of X_1, X_2, X_3 , and Y underlying the CC table; rather, it is the joint distribution of X_1, X_2, X_3 , and Y marginalized across missingness patterns. Generally, if the missing-data mechanism is nonignorable, we will need to specify a different reference table for each missingness pattern.

In our model-based approach, the two crucial issues to be addressed are: (1) how to estimate the set of reference tables using well-established principles of efficient estimation; and (2) how to perform the imputation once these estimates are obtained. Two methods of estimation will be compared in Section 3.4; in Section 3.3 we briefly discuss various alternatives for imputation.

3.3 Single, Multiple, and Mean Imputation

Once the reference tables have been estimated, distributions for each individual's missing variables given the observed ones have been completely specified. In theory, these distributions could be used to analytically calculate correct point and interval estimates for any quantities of interest. In practice, however, these calculations are usually intractable; some other procedure is needed. Filling in the missing values by imputation is an attractive alternative, because it creates a completed dataset, which can be analyzed by complete-data methods. Little (1986) summarizes the strengths and weaknesses of various imputation methods; we shall only comment on aspects relevant to the PES.

In current practice, each missing value is typically filled in by taking a single random draw from a distribution, thereby producing a simulated complete dataset, which is analyzed in the usual complete-data fashion. Interval estimates derived from this method will be artificially too precise, because they do not reflect the uncertainties of the imputation. One remedy for this, which is coming into use, is multiple imputation (Rubin 1987), in which each missing value is replaced by m random draws from the distribution. With moderate amounts of missing information, $m = 5$ draws are enough to produce efficient point estimates and adequate interval estimates. With rates of missing information that appear likely in the PES (typically 5 – 10 percent or less, judging from TARO), $m = 2$ draws will be perfectly adequate for essentially all purposes. In a large-scale survey like the PES, however, even a small number of multiple imputations may be computationally difficult to handle.

Since the estimates of interest in the PES are the match rates within poststrata, it is probably more important to accurately reflect the variability of imputation for Y than for X ; that is, it is probably more important to reflect uncertainty in overall undercount rates than uncertainty in the allocation of undercount to poststrata. Thus it may be possible to obtain adequate results by imputing a single set of X values, and then multiply imputing Y given X . Yet another possibility is to impute a single set of X values, and then impute the probability of match given X . This approach was used in TARO (Schenker 1988); it allows the imputed X 's and fractional Y 's to be treated like single imputations when estimating undercount rates.

Choosing an acceptable imputation procedure given a set of reference tables is the subject of ongoing research. It is hoped that the TARO approach of imputing a single value of X and then imputing $P(Y = 1 \mid X)$ will prove to be a useful compromise between the accuracy of multiple imputation and the computational ease of single imputation.

3.4 Models and Methods of Estimation

In this section, we present two alternative procedures for modeling the missing data and estimating the reference tables for imputation. The two procedures are the Ignorable Maximum-Likelihood (IML) method and a new Nonignorable Bayesian (NB) method that should be an improvement over IML if the missing data are not MAR.

3.4.1 The Ignorable Maximum-Likelihood Method

As mentioned previously, an ignorable imputation procedure needs to specify only a single reference table and apply it to all missingness patterns. One naive approach is to estimate this common reference table Θ by the cell proportions observed in the CC table. The resulting estimate $\hat{\Theta}_{CC}$ is asymptotically unbiased for Θ if the missing data are missing completely at random (MCAR), that is, if the probability of missingness for each item is completely independent of the data values, observed or missing. If the missing data are merely MAR, and not MCAR, then using $\hat{\Theta}_{CC}$ for imputation introduces biases into the data. Moreover, even when the data are MCAR, $\hat{\Theta}_{CC}$ is not efficient because it does not make use of all of the observed data to estimate Θ .

The IML method makes use of all the data, both in the CC table and in the supplementary tables, to estimate Θ . The estimated value $\hat{\Theta}_{IML}$ is chosen to maximize the likelihood ignoring the missing-data mechanism (Little and Rubin 1987, Section 5.3). In general, there is no closed form expression for $\hat{\Theta}_{IML}$; it must be obtained iteratively, for instance via the EM algorithm (Dempster, Laird and Rubin 1977; Little and Rubin 1987, Section 9.3).

The EM algorithm for contingency tables is easy to implement, and the resulting maximum likelihood estimate $\hat{\Theta}_{IML}$ is both efficient and consistent under the assumption of ignorability; thus this EM procedure for IML is attractive from both computational and theoretical perspectives. When the missing data are not MAR, however, the IML method will generally introduce biases. Since there are good reasons to believe that the missing data in the PES are not missing at random, we propose a new method of estimation that makes a different assumption.

3.4.2 Nonignorable Modeling and Nonuniqueness of the MLE

When the missing data are not MAR, it is no longer valid to ignore the missing-data mechanism; the fact that a data value is missing conveys information about its value. Hence, a model that reflects this dependence must include indicator variables for response, indicating whether data values were observed or missing. Consequently, a nonignorable model will generally estimate a separate reference table for each missingness pattern, or equivalently, an expanded reference table Θ with twice as many dimensions (*i.e.*, with an additional dimension for each missingness indicator).

Let $R = (R_1, R_2, R_3, R_Y)$ be indicator variables for whether X_1, X_2, X_3 , and Y are observed, respectively; for example, $R_1 = 1$ if X_1 is observed and $R_1 = 0$ if X_1 is missing. Consider the eight-dimensional contingency table formed by cross-classifying individuals by X, Y , and R , and now let Θ be the eight-dimensional table of cell probabilities for this expanded table.

Each individual in the survey belongs to a cell of the expanded table, but because some data are missing, we only observe certain margins of this table. Because R is fully observed, any margin involving only missingness indicators is fully observed, but a margin involving Y or one of the X 's might not be observed. For example, in the cross-section of the table with $R_1 = R_2 = R_3 = 1$ and $R_y = 0$, we can classify individuals by X_1 , X_2 , and X_3 , but not by Y ; therefore we observe only the marginal totals obtained by summing across Y .

The number of parameters in the fully saturated model for this table is $2^5 IJK - 1$, which is larger than the number of observed sufficient statistics; hence the maximum-likelihood estimate (MLE) for Θ is not uniquely determined. In order to obtain a unique estimate for Θ , one must impose additional structure.

One possible way to obtain a unique MLE is to build a log-linear model for the expanded contingency table, with some of the higher-order interactions set equal to zero (Little 1985; Fay 1986; Little and Rubin 1987, Section 11.6). We might try to set to zero those interactions that are not estimable from the data, but the formalization of this does not always work well in practice. For example, it may at first appear that the R_1 by X_1 interaction is not estimable, because the value of X_1 is never observed when $R_1 = 0$; however, the data may contain information about the R_1 by X_1 interaction indirectly through another variable, one that is observed for some individuals having $R_1 = 1$ and some having $R_1 = 0$. An example of a quantity that is truly inestimable from the data is $P(Y = 1 \mid X_1 = i, X_2 = j, X_3 = k, R_1 = R_2 = R_3 = 1, R_y = 0)$, but this does not correspond to any single interaction term in the log-linear model parameterization. (By "truly inestimable" we mean in Rubin's (1974) sense that the parameter's posterior distribution equals its prior distribution for all priors).

In a dataset with a complicated pattern of missingness, it is not easy to find a set of log-linear terms that, if set to zero, will yield a unique MLE for Θ . The minimum number of terms that must be set to zero to produce uniqueness is $2^5 IJK - 1$, the dimension of Θ , minus the number of observed sufficient statistics. Even if such a minimal set can be found, it is usually not unique, and one is faced with the task of deciding which set of terms should be excluded from the model. Rather than attempting to obtain a unique MLE by placing these kinds of prior restrictions on the log-linear model, we will instead use a Bayesian approach involving the use of a prior distribution.

3.4.3 A Nonignorable Bayesian Method

In the Bayesian paradigm, one expresses prior assumptions about the parameters formally through a prior distribution. For our situation, a proper unimodal prior, when combined with the observed-data likelihood, produces a posterior distribution for Θ that can yield a unique estimate; for example, we may take the posterior mode, $\hat{\Theta}_{NB}$, as our estimate of Θ . This method is attractive because it automatically allows precise estimation of those functions of Θ about which the data contain much information, while using the prior to select appropriate values for those quantities that are strictly inestimable from the data. If applied properly, this method will produce a nonignorable model that fits the data as well as any other model — it essentially maximizes the likelihood function, and yet is as consistent as possible with our beliefs about the nature of the missing-data mechanism as expressed in the prior distribution.

Sound scientific practice suggests that we should choose a prior distribution that favors simple structure (*i.e.*, small higher-order interactions) over complicated structure (*i.e.*, large higher-order interactions). If we choose a prior that assigns a low (but nonzero) *a priori* probability to the presence of higher-order interactions in the log-linear model, then we will be making assumptions that are similar in nature to the assumptions of the IML method — that

missing values are not radically different from their observed counterparts in their relationships with other observed variables – although in a smoother, more systematic fashion than the IML method does.

Following the notation of Bishop, Fienberg, and Holland (1975), consider the saturated log-linear model for the eight-way contingency table for R , X , and Y ,

$$\begin{aligned} \log \theta_{ijk\dots p} = & \mu + \mu_{1(i)} + \mu_{2(j)} + \dots + \mu_{8(p)} \\ & + \mu_{12(ij)} + \mu_{13(ik)} + \dots \\ & + \mu_{123\dots 8(ijk\dots p)}, \end{aligned} \tag{2}$$

where $\theta_{ijk\dots p}$ is the probability that an observation falls in cell $ijk\dots p$, and the μ 's are the one-way, two-way, three-way, and higher-order interactions. We propose the simple family of independent normal prior distributions

$$\begin{aligned} \mu_i &\sim N(0, \sigma^2) \\ \mu_{ij} &\sim N(0, \sigma^2/\tau) \\ \mu_{ijk} &\sim N(0, \sigma^2/\tau^2) \\ &\vdots \\ \mu_{ijk\dots p} &\sim N(0, \sigma^2/\tau^7), \end{aligned} \tag{3}$$

for some choice of $\sigma^2 > 0$ and $\tau > 1$. This prior distribution pulls the higher-order interactions toward zero, and hence pulls the estimate of Θ toward a more parsimonious or simpler model. We believe that this approach will produce estimates of Θ that are not too different from $\hat{\Theta}_{IML}$ when the missing data are truly MAR, but will be more robust than the IML method under departures from MAR. The only cases when IML will be superior occur when the missing data are MAR and strong higher-order interactions exist among the X 's and Y .

Leonard (1975) and Laird (1978) examined log-linear models with normal prior distributions on the μ terms for complete data; our situation is complicated by the fact that only certain margins of the eight-way table are observed. Finding the posterior mode $\hat{\Theta}_{NB}$ under this model is conceptually straightforward; the EM algorithm can be applied to the posterior distribution of Θ , just as to the likelihood function. The E-step remains the same; the M-step, however, poses some computational difficulties. The posterior distribution is nearly a ridge in high-dimensional space; it is very steep in certain directions, but nearly flat in others. The second-derivative matrix is nearly singular along this ridge; hence Newton-Raphson and other gradient methods for maximization will not work well. Difficulty arises as σ^2 becomes large, because the ridge becomes flat as $\sigma^2 \rightarrow \infty$ and a unique mode no longer exists. Difficulty also arises as the number of observations grows, because the posterior becomes very steep in certain directions and thus portions of the second-derivative matrix become very large. More work is needed to develop effective methods for finding or approximating $\hat{\Theta}_{NB}$.

3.4.4 A Numerical Example

We now present a simple numerical example and compare the results obtained from the IML and NB methods. For simplicity, we will only use a single dichotomous X variable (taking values 0 or 1) and match status Y .

If there were no missingness, the data could be fully cross-classified by X and Y and hence summarized in a single 2×2 contingency table. With four patterns of missingness, however, the data are summarized in a CC table and three supplementary tables (Figure 1).

The CC estimate $\hat{\Theta}_{CC}$ is simply the observed proportions in Table A. The IML estimate $\hat{\Theta}_{IML}$ is found iteratively via the EM algorithm; using $\hat{\Theta}_{CC}$ as the starting value, the algorithm converges in approximately four cycles. The NB estimate $\hat{\Theta}_{NB}$ was found using a prior distribution with $\sigma^2 = 10$ and $\tau = 3$. This means that the one-way terms are *a priori* normally distributed about zero with variance 10, so there is a 95 percent probability that the log-odds for each main effect lies inside the interval $(-4 \sqrt{10}, +4 \sqrt{10})$. The two-way terms have variance $10/3$, the three-ways have variance $10/9$, and the four-ways have variance $10/27$; this represents a moderate pulling of the higher-order terms toward the origin. (Finding $\hat{\Theta}_{NB}$ for varying values of σ^2 and τ proved difficult, because of the numerical instability of the particular maximization routine applied at each M-step.) The values of $\hat{\Theta}_{IML}$ and $\hat{\Theta}_{NB}$ are given in Figure 2. The expected imputations under these models are given in Figure 3, along with the expected imputations under $\hat{\Theta}_{CC}$ for comparison.

The differences between the imputation methods can be seen most clearly by comparing the expected imputations for Table D. Imputation using $\hat{\Theta}_{CC}$ simply reproduces the proportions observed in Table A. Imputation using $\hat{\Theta}_{IML}$ differs from imputation using $\hat{\Theta}_{CC}$ because Tables B and C, as well as Table A, contribute to the estimation of Θ and hence to the imputation for Table D.

Imputation using $\hat{\Theta}_{NB}$ is fundamentally different from imputation using $\hat{\Theta}_{CC}$ or $\hat{\Theta}_{IML}$ in that it assumes missingness is informative. From Table B, it surmises that missingness of Y is associated with $X = 0$. From Table C, it surmises that missingness of X is associated with $Y = 0$. It then combines this information in a smooth fashion to conclude that a larger proportion of the individuals who have both X and Y missing fall into the $(X = 0, Y = 0)$ category.

4. DISCUSSION

Our work is clearly at an early stage of development. Nevertheless, we feel that it has important potential applications, both specifically to the estimation of undercount using a PES, and generally to contingency table modeling when some data are missing. We conclude with two brief comments: first, on the need for continuing research on these procedures; and second, on the need to judge the relative propriety of models when devising an imputation procedure.

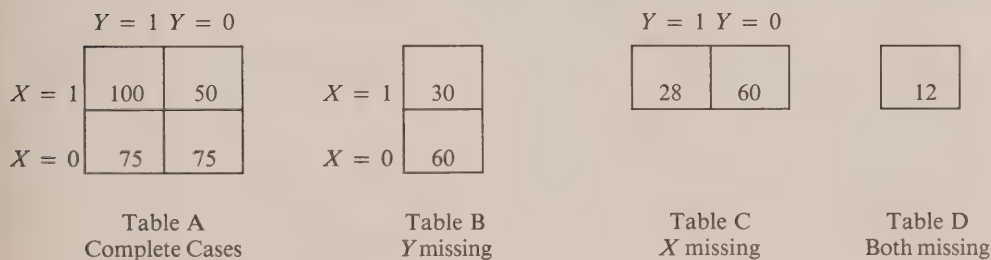


Figure 1. Observed Data

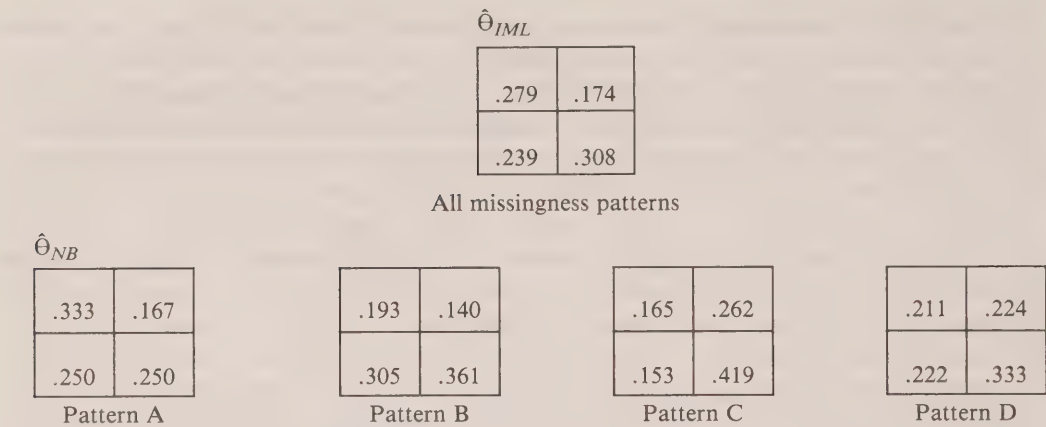


Figure 2. Reference Tables for Imputation

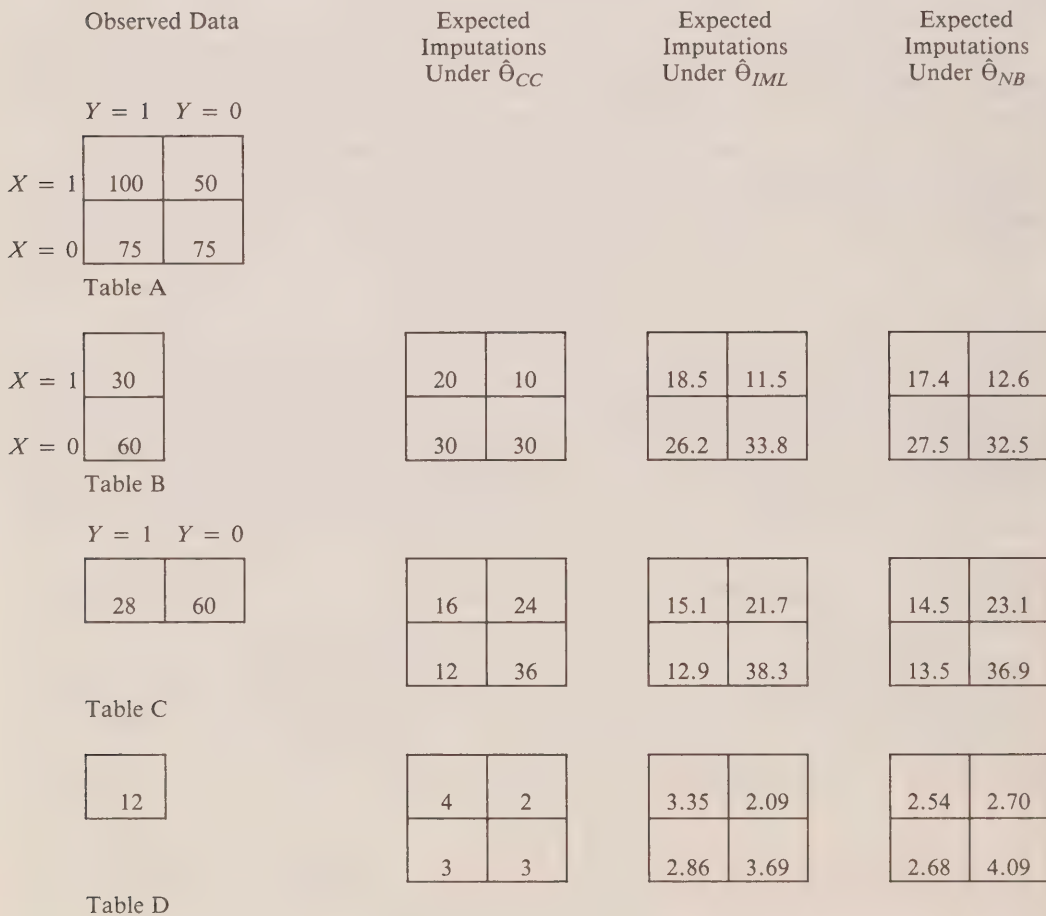


Figure 3. Expected Imputations Under $\hat{\Theta}_{CC}$, $\hat{\Theta}_{IML}$, $\hat{\Theta}_{NB}$

4.1 Continuing Research

Two kinds of research efforts are needed before our NB method can become broadly applicable. First, computationally-oriented research is needed to address the ridge-like posterior distribution. Alternatives to the mode, such as the posterior mean, are worth considering. Furthermore, measures of uncertainty should also be calculated, and considering the odd non-normal shape of the posterior, these may not be simple to summarize or compute. One strategy focuses directly on drawing multiple values of Θ from this posterior distribution without explicitly finding the posterior mode or the mean; these draws of Θ may be used to multiply impute the missing data.

Related to the issue of measuring uncertainty is the issue of performance in repeated sampling experiments. Although we believe our Bayesian approach is fully appropriate, it is important for broad application to evaluate the operating characteristics of this procedure in the wide range of circumstances to which it might be routinely applied. For example, how well does it work in realistic cases when, unknown to the data analyst, the missing data are MAR?

These topics will be the focus of a major continuing research effort.

4.2 The Need to Judge the Relative Propriety of Models

Considering the fully saturated model for (X, Y, R) with parameter Θ , any method of imputation, no matter how illogical, can be viewed as the correct procedure under some model. For example, consider imputation using $\hat{\Theta}_{CC}$ as the reference table for all missingness patterns. This posits conditional distributions for the missing data, given the observed data and R , about which there is no information in the observed values. Hence, coupling these distributions with the estimable distributions (the distributions of R and the observed data) implies an estimate for Θ , which maximizes the likelihood under the saturated model! It is not a very sensible answer, since it corresponds to the unique MLE under a model in which all sorts of conditional distributions given various missingness patterns R are equal to the conditional distributions given $R = (1, 1, \dots, 1)$; however, if we consider the likelihood function only, there is no reason to prefer any other maximum-likelihood estimate to this one.

Even stranger methods of imputation, such as "impute all missing values as zero," correspond to particular models with estimated Θ 's that are MLE's under the saturated model, but they violate good sense. Any sensible attempt to impute missing data values is based on the belief that two individuals with similar values of observed characteristics, and similar missingness patterns, are not radically different in those characteristics that are observed for one and missing for the other. Our NB method formalizes this notion of smoothness by specifying a contingency table model with small higher-order interactions.

Choosing one imputation procedure over another, then, cannot be done on maximum-likelihood-type principles alone, but must involve consideration of the propriety of the underlying prior specifications. This is not really a serious problem; sound statistical practice has always advocated the use of smooth or parsimonious models when less smooth models fit the data equally well. Consider fitting straight lines or polynomial curves through a collection of data points; simpler models are preferable to complicated ones on scientific grounds – the same issues arise in imputation. We believe that the model, given by (2) and (3), underlying our NB method, will be reasonable in many problems, just as linear regression is a reasonable tool in many problems.

ACKNOWLEDGEMENTS

This paper reports research undertaken primarily while Nathaniel Schenker was employed by the Statistical Research Division, Bureau of the Census, Washington, DC 20233, USA. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau. This research was supported in part by Joint Statistical Agreements 87-07 and 88-02 between the U.S. Bureau of the Census and Harvard University, and in part by the U.S. National Science Foundation under grant SES-88-05433, and represents a clarification and revision of Rubin, Schafer, and Schenker (1988). The authors wish to thank the two referees for their very helpful comments.

REFERENCES

- BAKER, S.G., and LAIRD, N.M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association*, 83, 62-69.
- BISHOP, Y.M.M., FIENBERG, S.E., and HOLLAND, P.W. (1975), *Discrete Multivariate Analysis*, Cambridge: MIT Press.
- DEMPSTER, A.P., LAIRD, N.M., and RUBIN, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39, 1-38.
- DIFFENDAL, G. (1988). The 1986 test of adjustment related operations in Central Los Angeles County. *Survey Methodology*, 14, 71-86.
- FAY, R.E. (1986). Causal models for patterns of nonresponse. *Journal of the American Statistical Association*, 81, 354-365.
- FAY, R.E., PASSEL, J.S., and ROBINSON, J.G. (1988). *The Coverage of Population in the 1980 Census*. 1980 Census of Population and Housing Evaluation and Research Report PHC80-E4, Washington: U.S. Government Printing Office.
- FREEDMAN, D.A., and NAVIDI, W.C. (1986). Regression models for adjusting the 1980 Census. *Statistical Science*, 1, 3-39.
- HOGAN, H., and WOLTER, K. (1988). Measuring accuracy in a Post Enumeration Survey. *Survey Methodology*, 14, 99-116.
- KROTKI, K.J. (1978). *Developments in Dual System Estimation of Population Size and Growth*, Edmonton: The University of Alberta Press.
- LAIRD, N.M. (1978). Empirical Bayes methods for two-way contingency tables. *Biometrika*, 65, 1, 581-590.
- LEONARD, T. (1975). Bayesian estimation methods for two-way contingency tables. *Journal of the Royal Statistical Society, Series B*, 37, 23-37.
- LITTLE, R.J.A. (1985). Nonresponse adjustments in longitudinal surveys: models for categorical data. *Bulletin of the International Statistical Institute*, 15, 1-15.
- LITTLE, R.J.A. (1986). Missing data in Census Bureau surveys. *Proceedings of the Second Annual Research Conference*, United States Bureau of the Census, 442-454.
- LITTLE, R.J.A., and RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*, New York: Wiley.
- MARKS, E.S., SELTZER, W., and KROTKI, K.J. (1974). *Population Growth Estimation*. New York: The Population Council.
- RUBIN, D.B. (1974). Characterizing the estimation of parameters in incomplete-data problems. *Journal of the American Statistical Association*, 69, 467-474.
- RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 3, 581-592.
- RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

- RUBIN, D.B., SCHAFER, J.L., and SCHENKER, N. (1988). Imputation strategies for estimating the undercount. *Proceedings of the Fourth Annual Research Conference*, United States Bureau of the Census, 151-159.
- SCHENKER, N. (1988). Handling missing data in coverage estimation, with application to the 1986 Test of Adjustment Related Operations. *Survey Methodology*, 14, 87-98.
- SCHENKER, N. (1989). The use of imputed probabilities for missing binary data. *Proceedings of the Fifth Annual Research Conference*, United States Bureau of the Census (forthcoming).
- WOLTER, K.M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, 81, 338-346.

The Sources of Census Undercount: Findings from the 1986 Los Angeles Test Census

DAVID J. FEIN and KIRSTEN K. WEST¹

ABSTRACT

This paper presents results from a study of the causes of census undercount for a hard-to-enumerate, largely Hispanic urban area. A framework for organizing the causes of undercount is offered, and various hypotheses about these causes are tested. The approach is distinctive for its attempt to quantify the sources of undercount and isolate problems of unique importance by controlling for other problems statistically.

KEY WORDS: Census; Undercount; Coverage improvement; Post enumeration survey.

1. INTRODUCTION

In the last decade or two the need to better understand the causes of undercount in the U.S. census has become pressing. As the census has become an increasingly important tool in governing the nation, conducting business, and monitoring social change (Citro and Cohen 1985; Clogg *et al.* 1986), public concern about the quality of census data has intensified. Much of this concern has arisen because it is perceived, with good foundation, that net census undercount disproportionately affects the economically disadvantaged members of society (Citro and Cohen 1985, ch. 5; Ericksen 1983). Representatives of the disadvantaged believe that as a result their constituents are being denied a fair share of public funds and political representation (Choldin 1987).

Assuming that an acceptable method could be found, one solution to the problem would be to correct the census for the bias due to differential undercount. In the fall of 1987, however, the Department of Commerce decided not to adjust the 1990 census but instead to concentrate on achieving a more complete enumeration (Ortner 1987).

Improving census coverage implies a need to understand the causes of census undercount better than ever before. Many special coverage improvement programs were implemented in the 1980 census, and these may have contributed to the achievement of historically low levels of overall net coverage error. In spite of such efforts, wide socioeconomic coverage differentials have persisted. In response, the Census Bureau has embarked on a broad research program to identify the causes of undercount, concentrating on population subgroups that are especially difficult to enumerate.

This paper presents results from a study of the causes of census undercount in a hard-to-enumerate, largely Hispanic area in Los Angeles. The approach is distinctive for its attempt to quantify the sources of undercount and isolate problems of unique importance by controlling for other problems statistically.

Though the putative inequities mentioned above result from net census coverage error (omissions less erroneous enumerations), to keep the analysis manageable only census omissions are investigated here. Omissions in the U.S. census deserve a higher position on the research agenda

¹ David J. Fein and Kirsten K. West, Undercount Research Staff, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C. 20233. The views expressed in this paper are those of the authors and do not necessarily reflect those of the Census Bureau.

because they are more numerous, vary more systematically with socioeconomic characteristics, and have been more politically controversial than erroneous inclusions.

The paper begins by describing a system for classifying the causes of undercount. Methods and results are presented next. A concluding discussion summarizes the implications for coverage improvement.

2. RESEARCH MODEL

The research model is presented in Figure 1. It represents undercount as a problem that occurs primarily at the household, rather than the individual, level. This specification is consistent with the basic sources of undercount in a census based on contacting each household rather than every individual in the population.

Three different household-level undercount problems are distinguished in the top margin of Figure 1: the omission of an entire household due to failure to enumerate a physical housing unit, the omission of an entire household in an enumerated housing unit, and the omission of only some members in a household where others are enumerated. Each of the three undercount problems can originate in census operations, in the society being enumerated, or in an interaction between operational and social system features. The following discussion is restricted to errors associated with the mailout/mailback methods used in the 1986 Los Angeles test census for a largely low income, Hispanic population.

2.1 Implementation of Census Operations

Operational difficulties during the census can cause the omission of housing units, of households in enumerated units, and of individuals in enumerated units. Occupied housing units can be missed because they are never added to the address lists or because they are on the lists but are erroneously deleted (U.S. General Accounting Office 1980). Given that a housing unit is correctly listed, all of the persons living in that unit may still be missed by the census due to misclassification of occupied units as vacant during nonresponse followup (U.S. Bureau of the Census 1987b; Ericksen 1983).

For questionnaires which households complete and mail back there are relatively few procedures for detecting missing persons. Procedures aimed at improving within household coverage include a question asking respondents if they were uncertain about including anyone and a clerical consistency check between a roster of household members requested at the beginning of the questionnaire and the number of persons for whom data are provided later on in the form (U.S. Bureau of the Census 1987b; Edson 1987). These procedures "cause" within household omission if they do not operate as intended due to errors in the administration of edit followup. Similarly, errors by enumerators during mail nonresponse followup may result in failure to add persons who should have been added.

Another important census operation is public information. Census publicity programs are designed to motivate mail response and reduce deliberate concealment by educating people about the uses of census data, the importance of complete reporting, and the confidentiality of census records. The extent to which such programs can reduce within household omission is unknown.

2.2 The Social System

At each stage of the census, data collection procedures come into contact with a social system which has many attributes that can impede enumeration. These attributes include unwillingness

Sources of Undercount for . . .

Missing Housing Units

Missing Households in Enumerated Units

Missing Persons in Enumerated Households

<ul style="list-style-type: none">• Address list omission• Erroneous deletion	<ul style="list-style-type: none">• Misclassification of occupied housing units as vacant	<ul style="list-style-type: none">• Failure of coverage edit procedures• Enumerator errors• Ineffective public information
<ul style="list-style-type: none">• Physical characteristics affecting unit visibility	<ul style="list-style-type: none">• Factors affecting household visibility• Factors leading to refusal	<ul style="list-style-type: none">• Factors causing unwillingness to report• Factors causing respondent definitional error

Census Implementation

Social System

Figure 1. Research Model

to report some or all household members, inability to report in a manner consistent with census definitions, and low "social visibility" of household members or the housing units in which they live. (Social visibility is the degree to which household members and housing units possess characteristics which make them perceptible to outsiders.)

The most important social system factors causing housing unit omission are those affecting the social visibility of units. Some kinds of units are easier to find and more likely to appear on commercial address lists than others. Social system sources of omission for households in enumerated units include factors depressing the visibility of household members and refusal to report.

All three broad sets of social system causes are implicated in within household omission: unwillingness to report, definitional problems, and the differential social visibility of household members. Willingness to report can be approached by considering the perceived costs and benefits of reporting for respondents (Dillman 1978). There has been much discussion of the perceived costs of census reporting. People may fear that disclosure of adult males will jeopardize welfare eligibility, that persons illegally in the country will be deported, that reporting more persons than allowed by a lease will prompt landlord troubles, and that police will be informed of the whereabouts of lawbreakers (Bailar and Martin 1987). Such fears may cause noncompliance when there is disbelief in the Census Bureau's promise of confidentiality.

The sources of definitional error are quite different from those of concealment. Definitional errors arise in the complexities of household living arrangements, as conditioned by respondents' abilities to understand and apply census enumeration and residence rules (Hainer *et al.* 1988).

Having mentioned some of the major sources of undercount, we will now examine the extent to which they occurred during the 1986 Los Angeles test census.

3. METHODS

3.1 Data Sources

This study takes an intensive look at undercount in a March 1986 test census conducted in the northern half of Los Angeles County. The population was low income and largely Hispanic. Nearly two-thirds (65%) of the heads of households enumerated in the census were of Spanish origin and 13% were Asian. Residences in this part of Los Angeles were largely single family dwellings (73%) and small apartment buildings (15%). Owners lived in half (51%) of the occupied units, in contrast with nearly two thirds (65%) of all occupied units nationwide (U.S. Bureau of the Census 1987a: 106, table 18; U.S. Bureau of the Census 1987c: 712, table 1285).

The data analyzed are from the 1986 Los Angeles test census itself; the Post Enumeration Survey, or PES, conducted to measure test census coverage; and a special followup to the PES—the Causes of Undercount Survey. The census enumerated 109,900 housing units and was intended primarily as a test of planned 1990 census operations.

The Post Enumeration Survey (PES) was one of these operations. The purpose of the PES, conducted in July 1986, was to identify census omissions and erroneous enumerations (Diffendal 1988). It did this by attempting to match PES to census records. When a PES person's record was found in the census it was termed "matched"; otherwise the person was considered "nonmatched".

Three kinds of PES households are distinguished here, depending on whether all, some, or none of their members were matched to the census. "Complete match" households contain only persons in the PES who were matched to persons in the census. "Partial nonmatch"

households contain at least one person who could not be matched and at least one person who was matched to the census. “Total nonmatch” households include only persons who could not be matched to the census.

These three household types are distinguished to allow examination of problems associated with housing unit omission, omission of entire households in enumerated units, and omission of persons from households that were partially enumerated. Completely matched households are included for reference purposes, to represent households correctly enumerated in the census.

A special followup survey – the Causes of Undercount Survey – was conducted in November 1987 to obtain additional information needed to compare these household types. The survey obtained information on census characteristics for nonmatched persons, as well as some new household and housing unit data not available on the census or PES files.

The entire partial nonmatch stratum and nearly all households in the total nonmatch stratum were selected for reinterview. Eight total nonmatch households had to be omitted because several items needed to reinterview them were missing. Households in the complete match stratum were subsampled to reduce survey costs.

The distribution of the 966 completed Causes of Undercount Survey interviews by household type is shown in the right-most column of Table 1. This table also gives the unweighted numbers for all 5814 PES households and the 1420 cases in the Causes of Undercount Survey sample. The overall response rate for the survey was 68%, reflecting considerable success in locating households in a transitory urban area despite the 16 months intervening between the survey and the PES.

3.2 Analysis Plan

There are several parts to the analysis. PES total nonmatch households are examined first. Two sets of comparisons are made: 1) of missed housing units with enumerated housing units and 2) of missed households in enumerated units with enumerated households. Missed housing units were expected to contain a higher percentage of clustered housing units and unusual unit types and locations than enumerated units. Missed households in enumerated housing units were expected to be smaller, contain adults who were less frequently at home, and move more often than enumerated households. Most of the explanatory variables for housing unit and household omission were obtained either from the census Address Control File or from the PES matched file, and thus are available for all 193 total nonmatch households in the sample.

Table 1
Numbers of Households in the PES and Causes of Undercount Survey Sample,
and Numbers of Completed Interviews, by Household Type.

Household Type	Post Enumeration Survey	Causes of Undercount Survey	
		Sample	Completed Interviews
Complete Match	4,871	489	382
Partial Nonmatch	738	738	484
Total Nonmatch	205	193	100
All Types	5,814	1,420	966

The second part of the analysis compares partial nonmatch with complete match households to identify factors responsible for within-household omission. Two sets of explanatory factors are distinguished, those indicating inadvertent or "definitional" errors and those representing reasons for deliberate concealment. Indicators for definitional errors include large size and complex composition of households, poorly-spoken English and educational deficits. Concealment indicators include presence of recent immigrants, welfare reciprocity, crowded housing, and disbelief in census confidentiality. It was hypothesized that partial nonmatch households would score higher on the definitional and concealment indicators than would complete match households.

The analysis begins with bivariate relationships between each of the explanatory factors and partial omission and then considers multivariate relationships. The source for many of these indicators was the Causes of Undercount Survey; hence, only data from interviewed households are used.

In the final part of the analysis, characteristics of four types of individuals are compared: persons *matched* in complete match and partial nonmatch households, and those *nonmatched* in partial and total nonmatch households. Characteristics compared include age, sex, education, relationship to the household head, and citizenship status.

Bivariate percentages are based on weighted data to compensate for the PES and Causes of Undercount Survey sampling designs, though tests for differences between these percentages used unweighted numbers. Unweighted data were used to estimate parameters of log-linear models. The effects of the PES sampling design on estimates for the final models were evaluated by adding in all two-way interactions which included the PES stratification variable. This adjustment did not greatly change the results; thus, the estimates presented here do not include the stratification variable. Because the second stage of PES sampling entailed cluster sampling of households in census blocks, the standard errors calculated are likely to underestimate the true sampling errors: they are presented only as rough guides to the significance of parameters.

4. FINDINGS

4.1 Total Nonmatch Households

Table 2 shows the final status assigned in the census to PES total nonmatch households for cases sent and not sent to nonresponse followup. Of the 193 total nonmatch cases 97, or 50%, never appeared on the census address lists. Thus, housing unit omission appears to explain why the PES could not find anyone in these households in the census.

The remaining 96 cases did appear on the census address lists. What caused these households to be missed? The explanation is probably that most of these units were census closeout interviews, where a landlord or neighbor provided only an estimate of the total number of persons in the household and not detailed information for individuals. This hunch is supported by the finding that of the 44 cases the census classified as occupied, population counts for 37 were "goldplated". This means that the final count accepted for these households was not obtained in the usual manner by allowing the FOSDIC (Film Optical Device for Input to Computers) machines to count persons. Instead, goldplating involved accepting a total count for the household entered on the questionnaire in the field. This is likely an indication that the household was a closeout case.

Thus, the census really did not miss most of these 44 households entirely, though when it came time for PES matching, there were no individual census person records to be matched.

Table 2
Final Status Assigned in the Census to PES Total
Nonmatch Households By Nonresponse Followup Status:
Numbers of Units^a

Final Status of Unit in Census	Sent to Nonresponse Followup?		
	No	Yes	Total
Omitted from the Census Address Lists	97	0	97
Included in the Census Address Lists	4	92	96
Occupied, Direct Accept ^b	1	6	7
Occupied, Gold-plated ^c	2	35	37
Vacant, Direct Accept	1	34	35
Vacant, Gold-plated	0	17	17
All Units	101	92	193

Notes: ^a N's are unweighted.
^b Direct Accept: FOSDIC person count accepted.
^c Gold-plated: Field counts accepted instead of FOSDIC.

An allowance is made for these cases in the dual system estimation method. Nevertheless, it still is true that these households were not directly enumerated.

To summarize, 50% of the PES total nonmatch households were in units which appeared to have been entirely omitted. Of the households living in units which were enumerated, 54% had been classified as vacant, possibly erroneously, and 46% had been found to be occupied. Of the total nonmatch households classified as occupied in the census, up to 84% may have been enumerated in closeout interviews.

Figure 2 compares some physical characteristics of units left off the census address lists (light bars) with units that were not left off the lists (dark bars). The top set of bars represents the basic types of housing units. Attached single family homes, such as duplexes, appear to have been a major problem in the L.A. test census. Thirty-four percent (34%) of the missed units fell into this category, in contrast to only 8% of enumerated units. Missed units were less likely than enumerated units to be detached single family homes or apartments in large buildings, suggesting that the census was more successful at finding such units.

Whether or not an interview was completed, Causes of Undercount Survey interviewers were asked to record when units they visited fit any of several "unusual unit" categories listed on the front of their questionnaires. The bottom half of Figure 2 shows that the interviewers identified a higher percentage of unusual units among units that were missing from the census address lists, 28%, than among units that were included, 7%. Unit types found to be particular problems were abandoned-looking buildings and secondary units on a lot.

Physical characteristics of units thus do appear to affect their visibility during census address list development. What might cause households to be missed in units that were enumerated?

Households may be more easily missed if they are small and mobile. Figure 3 compares characteristics of total nonmatch households in enumerated units with a combined group of complete match and partial nonmatch households - that is, households which were enumerated. Households missed in the test census (light bars) were on average considerably smaller than those where some or all members were counted (dark bars). Whereas 53% of the total nonmatch households in enumerated units had one or two members, only 35% of the enumerated households were this small.

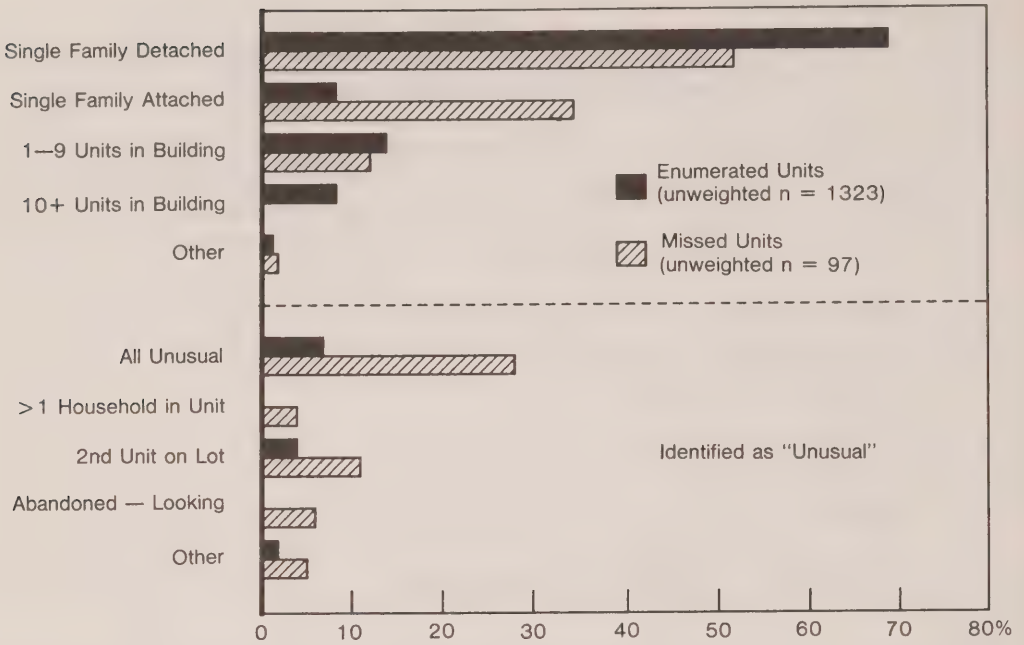


Figure 2. Physical Characteristics of Enumerated and Missed Housing Units (Weighted Percentages)

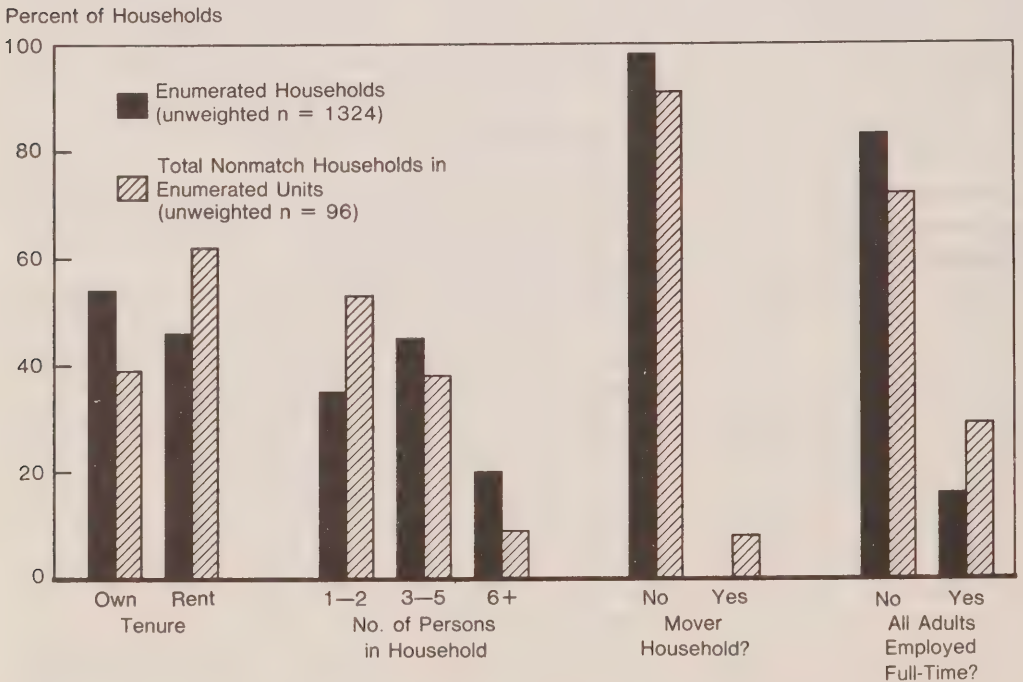


Figure 3. Characteristics of Enumerated Households and Total Nonmatch Households in Enumerated Units (Weighted Percentages)

Indicators of the propensity to move include home ownership and actual household mobility in the four months between the census and the PES. Households missed in the census were more likely to be renters and movers (61% and 8%, respectively) than were enumerated households (46% and 0%, respectively). The percentage of households in which all adults were employed full-time in March 1986 was greater by 12% for omitted households than for enumerated households, though the number of interviews for omitted households was too small for this difference to be statistically significant.

These results support the hypothesis that missed housing units and households missed in enumerated units possess attributes which reduce their visibility during a census.

4.2 Partial Nonmatch Households

From total nonmatch households, the focus shifts to the factors associated with partial household omission. In this phase of the analysis, 484 partial nonmatch households were compared with 331 complete match households. Single person households were excluded from the 382 complete match households in the Causes of Undercount Survey sample, since they were not at risk of partial omission.

Two different sets of explanatory factors were considered. The first represents household characteristics thought to be associated with definitional errors, described earlier as errors resulting from inconsistencies between household membership as understood by the Census Bureau and by census respondents. The second set of indicators represents factors thought to be associated with the deliberate concealment of household members.

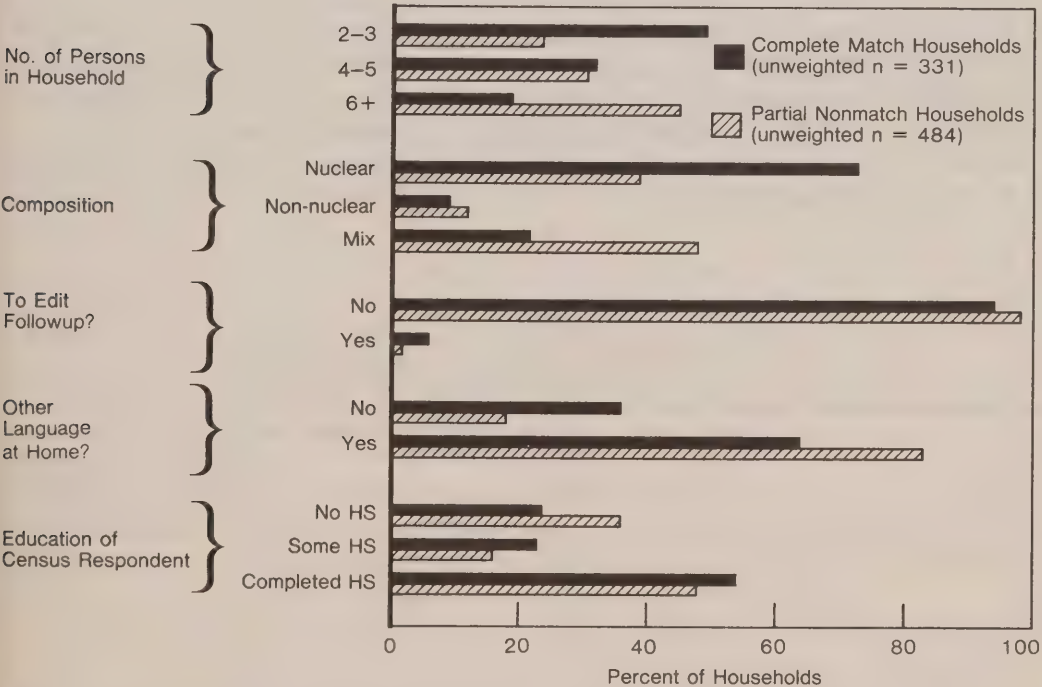


Figure 4. Definitional Error Indicators for Partial Household Omission: Households with 2+ Persons (Weighted Percentages)

Definitional Errors

Indicators for definitional errors include household size and composition, English language ability, census respondent's education, and edit followup status. Larger households, those containing more distant relatives and persons unrelated to the household head, those speaking a language other than English at home, those where the census respondent's education was low, and households not sent to edit followup were all expected to be at greater risk of definitional errors.

Figure 4 supports these hypotheses. It shows that partial nonmatch households (light bars) were considerably larger than complete match households (dark bars): 45% of the partial nonmatch households but only 19% of the complete match households contained six or more members. Whereas 40% of the partial nonmatch households contained only nuclear relatives of the household head, fully 72% of the complete match households were nuclear. Partial nonmatch households were less likely to have been sent to edit follow-up by a slight, but statistically significant, amount. Partial nonmatch households were more likely to speak a language other than English at home (83%) than were complete match households (64%). Finally, census respondents from partial nonmatch households had less formal education than those from complete match households: 36% of the census respondents from partial nonmatch households had not attended high school, in contrast with 24% of the respondents from complete match households.

Log-linear models were fitted to see whether these differences persisted at the multivariate level. The dependent variable in these models was partial household omission, with complete match households coded as 0 and partial nonmatch households coded as 1. Interactions between partial omission and each of the independent variables in Figure 4 were tested in a series of nested models. All two-way interactions among independent variables were included in each model as controls.

In the multivariate analysis, significant interactions with partial omission were found for all definitional error indicators except census respondent's education. Table 3 presents the chi square (Wald) statistics associated with the final definitional model, which excludes census respondent's education. Significant interactions of household size with composition and language other than English were also detected. Parameter estimates in Table 4 show the effects to be in the directions expected. Estimates for standardized parameters, obtained by dividing

Table 3
Chi Square Statistics For Testing Two-Way Interactions
in the Final Definitional Error Model^a

Variables	Interactions with . . .			
	Size	Composition	Edit Followup	Language at Home
Partial Omission	38.1**	42.3**	6.3*	5.2*
Size	-	112.0**	.9	50.0**
Composition	-	-	1.6	1.3
Edit Followup	-	-	-	1.0

** : p < .01
* : p < .05
^a Log Likelihood X² = 42.2, df = 45, p = .5922.

Table 4
Parameter Estimates for Interactions Between Definitional Error Indicators and
Partial Household Omission in the Final Model

Marginals with Partial Nonmatch Household and . . .	Parameter Estimate	Standard Error	Standardized Parameter Estimate
Household Size:			
2-3 Persons	-.34	.06	-5.7
4-5 Persons	-.02	.05	-.4
Composition:			
All nuclear	-.36	.06	-6.0
All non-nuclear	.22	.09	2.4
Edit Followup Status			
Not sent	.25	.10	2.5
Other Language at Home?			
Yes	.10	.05	2.0

parameter estimates by their standard errors, indicate that the effects of size and composition are about the same in magnitude and that both are larger than the effects of edit followup and language spoken at home.

Concealment Indicators

Factors hypothesized to cause concealment of household members by census respondents include: fear that persons illegally in the country would be deported, fear that disclosure of adult males would jeopardize welfare aid, and concern that reporting more persons than allowed by a lease would bring landlord troubles. Indicators for these factors were, respectively, whether the household contained recent immigrants, defined as persons entering the country in or after 1980; whether anyone in the household was receiving welfare during the census month; and the average number of persons per room in the household. Nonresponse to the census mailout was also included as a general indicator of failure to perceive positive benefits from responding to the census. Finally, belief in census confidentiality was included to see whether it helped to reduce fears resulting in concealment.

Figure 5 shows that all of these indicators were related to partial omission at the bivariate level. For example, recent immigrants were present in 26% of the partial nonmatch households (light bars), but only 12% of the complete match households (dark bars). Whereas 24% of the partial nonmatch households reported receiving welfare, only 15% of the complete match households did so. Partial nonmatch households were considerably more likely to exhibit crowding: 63% contained more than one person per room, in contrast to only 34% of the complete match households. Partial nonmatch households were also somewhat less likely than complete match households to have returned their census questionnaires by mail or to believe in census confidentiality.

Again, loglinear models were fitted, with partial omission as the dependent variable and the concealment indicators as independent variables. All two-way interactions with household size were included as controls, since other things being equal, larger households would be more likely to exhibit crowding and contain recent immigrants than small ones.

This time, two variables did not survive preliminary testing: mail nonresponse and belief in census confidentiality. Before completely dropping the confidentiality variable, tests were performed to see if interactions of partial omission with presence of immigrants, welfare reciprocity, and crowding depended on belief or disbelief in confidentiality. Belief in confidentiality was not found to affect these relationships.

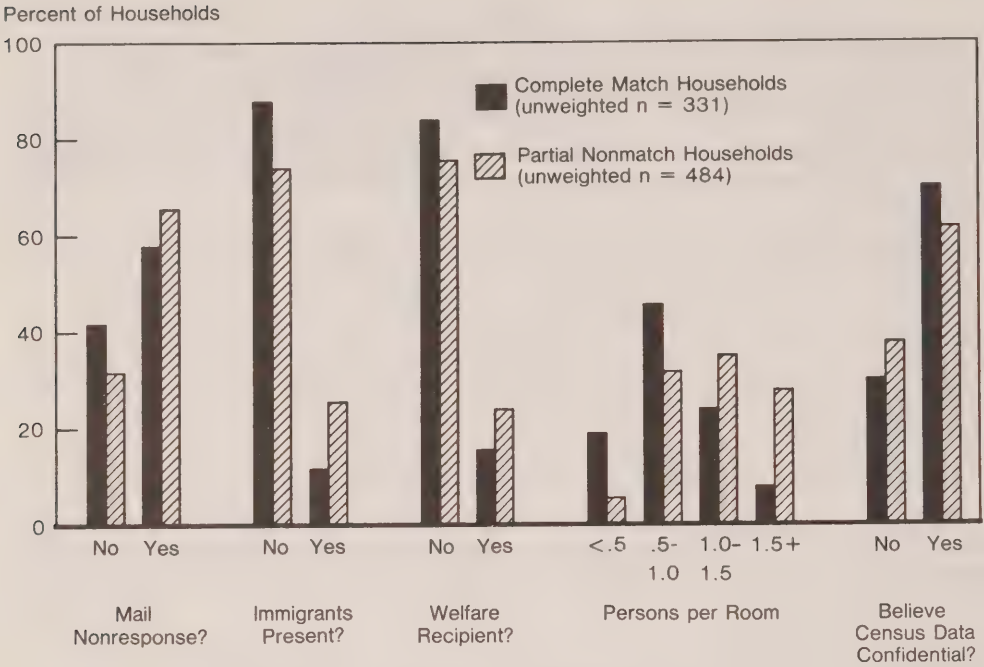


Figure 5. Concealment Indicators for Partial Household Omission: Households with 2+ Persons (Weighted Percentages)

Table 5
Chi Square Statistics For Testing Two-Way Interactions
in the Final Concealment Model^a

Variables	Interactions with . . .			
	Size	Immigrants	Welfare Assistance	Crowding
Partial Omission	2.9	11.3**	10.1**	16.7**
Size	-	.2	7.5*	221.7**
Recent Immigrants	-	-	1.6	30.0**
Welfare Assistance	-	-	-	5.4

** : p < .01
* : p < .05
^a Log Likelihood X² = 103.8, df = 150, p = .9985.

Table 5 shows that three of the remaining concealment variables immigrants, welfare, and crowding interacted significantly with partial household omission in a model which included all two-way interactions with size and all two-way interactions among independent variables. Standardized parameter estimates (see Table 6) suggest effects of roughly equal magnitude for the three indicators.

It is noteworthy that the relationship between partial omission and size vanished when crowding was included (see Table 5), suggesting that the effects of size were due to its association with crowding rather than scale alone. Crowding was also strongly associated with the presence of recent immigrants.

4.2 Person Characteristics

For the final part of the analysis of individual-level characteristics associated with undercount, four kinds of persons were compared: persons the census counted in complete match and partial nonmatch households, and persons the census missed in partial and total nonmatch households.

Figure 6 shows differences between the percentages in 10 year age groups for persons in complete match households and each of the three other groups. It shows an excess in the 20-29 year old group for persons missed in partial and total nonmatch households relative to persons in complete match households. There is also evidence of an excess in the 20-29 year age groups for persons who were enumerated in partial nonmatch households.

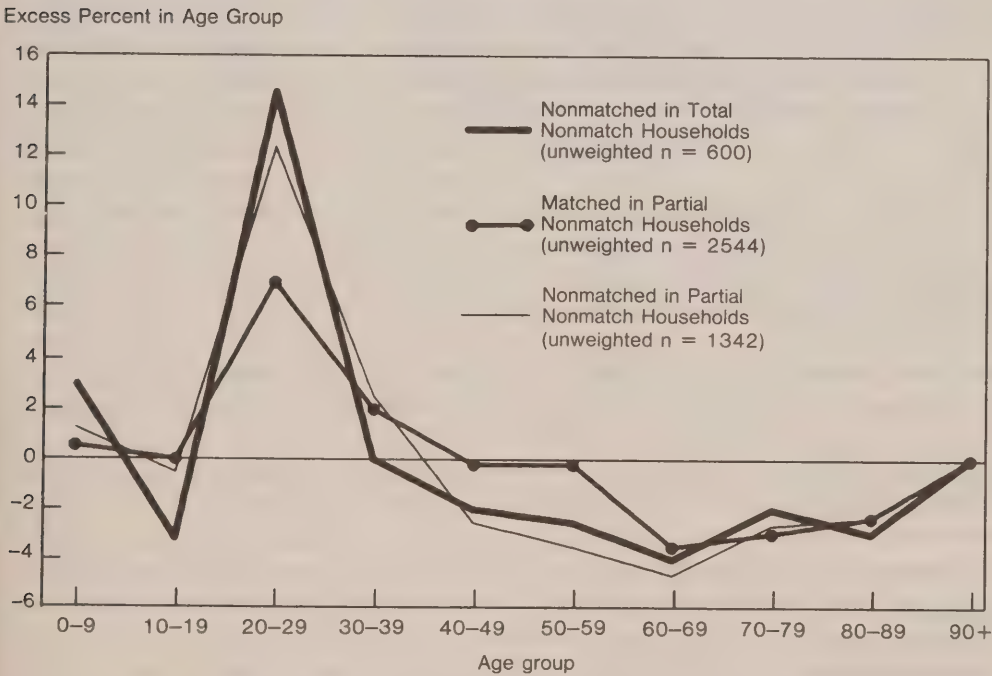


Figure 6. Excess Weighted Percentage in Age Group Relative to Persons in Complete Match Households

Table 6

Parameter Estimates for Interactions Between Concealment Indicators and
Partial Household Omission in the Final Concealment Model

Marginals with Partial Nonmatch Household and . . .	Parameter Estimate	Standard Error	Standardized Parameter Estimate
Recent Immigrants: Immigrants Present	.19	.06	3.2
Welfare Reciprocity: Receiving Aid	.17	.05	3.4
Crowding: < .5 Persons/Room	-.49	.13	-3.8
.5-1.0 Persons/Room	-.01	.08	-.1
1.0-1.5 Persons/Room	.08	.08	1.0

Table 7

Percentage Distributions for Characteristics of Individuals by
PES Match Status and Household Type

Characteristic	PES Match Status			
	Matched in		Nonmatched in	
	Complete Match HHs	Partial Nonmatch HHs	Partial Nonmatch HHs	Total Nonmatch HHs
Sex				
Male	46.2%	50.6%	54.2%	48.2%
Female	53.8	49.4	45.9	51.8
Unweighted <i>n</i>	1667	2564	1324	582
Education				
No Formal Education	10.2	10.9	17.0	14.3
Less than High School	30.7	34.4	27.2	37.5
Some High School	20.5	20.6	19.5	19.5
High School Graduate	38.6	34.1	36.4	28.8
Unweighted <i>n</i>	1197	1560	599	315
Relationship to Head				
Nuclear Relative	86.1	83.2	63.6	85.9
Non-nuclear Relative	11.3	12.6	25.4	7.9
Non-relative	2.6	4.2	11.0	7.0
Unweighted <i>n</i>	1659	2560	1359	590
Citizenship				
Citizen Since Birth	66.2	53.5	52.6	50.4
Naturalized Citizen	9.2	9.5	6.4	6.4
Noncitizen	24.6	37.0	41.0	43.2
Unweighted <i>n</i>	1223	1567	612	316

Persons missed by the census in partial nonmatch households were slightly more likely than persons in complete match households to be males and have no formal schooling, and less likely to be citizens or close relatives of the household head (Table 7). Persons missed by the census in total nonmatch households were also slightly more likely to be noncitizens and lower in education than persons in complete match households, but displayed no differences in sex and relationship to household head. Thus, on the whole, persons missed in partial nonmatch households differed from those in complete match households in more ways than did persons missed in total nonmatch households.

In addition to biasing more census characteristics, partial household omission caused the omission of many more persons than did total household omission. Two thirds (67%) of all PES nonmatch cases were in partial nonmatch households and only one third were in total nonmatch households. Fully 82% of all PES omissions were found in housing units the census enumerated and only 18% were in missed units.

5. DISCUSSION

The findings reported here support evidence from more qualitative studies that partial household omission is the most serious undercount problem in hard-to-enumerate urban areas of the United States today. As compared with total household omission, partial omission in the Los Angeles test census accounted for twice as many missing persons, reflected more intractable sources of error, and biased more individual-level census characteristics.

The chief problems identified for total household omission were failure to include certain types of housing units in the census address lists and misclassifying occupied units as vacant. Housing units especially at risk of misclassification as vacant were those with households which were small and mobile and those in which all adults were working full-time. Experience with coverage improvement programs at the Census Bureau suggests that further reductions in housing unit omission may be possible. Such programs were responsible for adding about 10% of the units enumerated in Los Angeles. The Bureau adopted special prec canvassing procedures in the test census to find units in large multi-unit structures. Considerable success in reducing this source of error in the test census is evident in Figure 2: none of the apartment units missed were in large buildings.

The misclassification of occupied units as vacant will be more difficult to remedy. Allowing nonresponse enumerators more time per unit and improved training for certain kinds of problem households may help somewhat. Coupling these efforts with special callback procedures for smaller and more transient households and those whose members are rarely at home would also help.

It is clear that improvements at the margin of what is already a largely successful census operation will be expensive. Keyfitz (1979) and others have observed that the incremental costs from adding persons to the count soar as coverage approaches 100%. Programmatic innovations to reduce the errors observed in the 1986 test census would add to the \$2.6 billion cost projected for the 1990 census, since the methodology to be used in urban areas will be very similar to the L.A. test census.

Within-household errors will be even more difficult to address than total household omissions. The Bureau must redouble its efforts to understand the complex living arrangements and cognitive and/or cultural factors that condition how people perceive household membership. The findings reported here suggest that further efforts targeted to respondents for whom English is not a native tongue, and households containing persons only distantly related to each other may help to reduce definitional errors.

However, in light of the considerable research already performed to improve the design of the census questionnaire and the complex enumeration and residence rules to which the Bureau is bound by statute and tradition, further reductions in definitional error will require extraordinary efforts. Definitional errors are deeply embedded in cultural differences and educational deficits among hard-to-enumerate groups.

Within-household omission also was found to be strongly related to the presence of immigrants, welfare reciprocity, and crowding. That a PES-based study could detect such effects suggests that the PES succeeded in counting many persons whose presence had been concealed in the census. Some of the effects of the so-called concealment variables may be due to uncontrolled factors other than concealment, but the persistence of relationships even after household composition was added in a final log-linear model (not shown) suggests that the PES really did detect some persons who were concealed in the census. Thus, there appears to be a continuum from households that are highly resistant to enumeration to those which are less resistant, and for the latter more intensive methods like those used in the PES may be effective.

The social conditions underlying the most resistant forms of concealment present the most difficult problems for the Census Bureau. Public information programs attempting to convince people that the census is important and that census data will be kept confidential were not very effective for the hard- to-enumerate population in the Los Angeles test census, as reported by Moore and McDonald (1987), though these programs may work better under real decennial census conditions. The minimal role found for belief in census confidentiality, either in its own right or in mediating between household circumstances and concealment, suggests that the relationship between attitudes and census response behavior is not a simple one.

The findings reported here should not be generalized uncritically to the sources of undercount expected to affect urban areas in the 1990 Census. Because the data are based on a test census, errors may reflect inexperience with experimental procedures or failure to convince respondents (and census workers) that the project was as serious as the decennial census. Further, to the degree that Los Angeles is unlike other major urban areas, it may experience unique census-taking problems. For example, Los Angeles is thought to be home to more "illegal aliens" than any other major city (Heer and Passel 1987).

On the other hand, the net undercount rate for Los Angeles in 1980 was quite similar to the rates for other major cities, as measured in the 1980 Post Enumeration Program (Fay *et al.* 1988). Thus, what they lack in illegal aliens, these cities may make up in other hard-to-enumerate groups. Further research is needed to assess the degree to which causes of undercount differ by race, ethnicity, and other social characteristics.

It is encouraging that the causes of undercount identified in this Post Enumeration Survey-based study were reasonably consistent with more qualitative reports by ethnographers and focus groups. Also, the PES estimates for undercount from the Los Angeles test census are believed to be of high quality (Hogan and Wolter 1988). For these reasons, extension of the PES-based methodology developed in this paper to other urban (and nonurban) areas is recommended.

On the social system side, further research on how rationally people weigh the costs and benefits of responding to censuses and surveys would help to weigh the potential for improving census coverage through the Census Bureau's public information and community action programs. Better indicators for household-level reasons for concealment are also needed. Examining specific assistance programs would help to confirm the effects of welfare participation on census coverage, since not all aid would be imperiled by revealing true household-composition.

Improved measurement of the sources of undercount arising in census operations is also needed. If data from census quality control programs were combined with PES matching results, error sources could be identified with greater precision.

ACKNOWLEDGEMENTS

We would like to thank the following persons for their assistance in this research: Irwin Anolik, Miriam Balutis, Gregg Diffendal, Chris Dyke, Sue Finnegan, Howard Hogan, Jan Jaworski, Pete Long, and Lynn Weidman. Betsy Martin, Jim O'Brien, and two anonymous reviewers provided valuable comments on earlier versions of this paper.

REFERENCES

- BAILAR, B., and MARTIN, E. (1987). Report on Meetings in Los Angeles, Chicago and Denver. Unpublished Census Bureau memorandum.
- CHOLDIN, H. (1987). Science and Scientists in the 1980 Census Lawsuits. Paper presented at the May 1987 meeting of the Population Association of America, Chicago.
- CLOGG, C.C., MASSAGLI, M.P., and ELIASON, S.R. (1986). Population undercount as an issue in social research, *Proceedings of the Second Annual Research Conference*. United States Bureau of the Census, Washington, D.C., 335-343.
- CITRO, C.F., and COHEN, M.L. (eds.) (1985). *The Bicentennial Census: New Directions for Methodology in 1990*, Panel on Decennial Census Methodology, National Research Council, Washington, D.C.: National Academy Press.
- DIFFENDAL, G. (1988). The 1986 test of adjustment related operations in Central Los Angeles County. *Survey Methodology* 14, 71-86.
- DILLMAN, D.A. (1978). *Mail and Telephone Surveys: The Total Design Method*. New York: John Wiley and Sons.
- EDSON, R.G. (1987). Preliminary coverage improvement results from tests for the 1990 Census. Paper presented at the August 1987 meeting of the American Statistical Association, San Francisco.
- ERICKSEN, E.P. (1983). Affidavit, Mario Cuomo, *et al.* vs. Malcolm Baldrige *et al.*, U.S. District Court, Southern District of New York, 80 Civ. 4550 (JES).
- FAY, R.E., PASSEL, J.S., and ROBINSON, J.G. (1988). The coverage of population in the 1980 Census. *Evaluation and Research Reports. 1980 Census of Population and Housing PHC80E4*, Washington, D.C.
- HAINER, P., HINES, C., MARTIN, E., and SHAPIRO G.M. (1988). Research on improving coverage in household surveys. *Proceedings of the Fourth Annual Research Conference*. United States Bureau of the Census, Washington, D.C., 513-539.
- HEER, D.M., and PASSEL, J.S. (1987). Comparison of two methods for estimating the number of undocumented Mexican adults in Los Angeles County. *International Migration Review*, 21(4), 1446-1473.
- HOGAN, H., and WOLTER, K. (1988). Measuring accuracy in a Post-Enumeration Survey. *Survey Methodology* 14, 99-116.
- KEYFITZ, N. (1979). Information and allocation: Two uses of the 1980 Census. *The American Statistician*, 33(2), 45-56.
- MOORE, J.C., and McDONALD, S.-K. (1987). The Census community awareness program: an evaluation of the potential and actual effectiveness of CCAP based on evidence from the 1986 Los Angeles Census Test. Unpublished Census Bureau report.

- ORTNER, R. (1987). Statement. *United States Department of Commerce News*, October 30, 1987.
- U.S. BUREAU OF THE CENSUS (1960). *The Post-Enumeration Survey: 1950*. Technical Paper No. 4, Washington, D.C.
- U.S. BUREAU OF THE CENSUS (1987a). 1986 Test Census, Central Los Angeles County, California. *General Population and Housing Statistics*, TC86-1, Washington, D.C.
- U.S. BUREAU OF THE CENSUS (1987b). Programs to improve coverage in the 1980 Census. *Evaluation and Research Reports. 1980 Census of Population and Housing*, PHC80-E3, Washington, D.C.
- U.S. BUREAU OF THE CENSUS (1987c). *Statistical Abstract of the United States: 1987*, (106th edition), Washington, D.C.
- U.S. GENERAL ACCOUNTING OFFICE (1980). *Problems in Developing the 1980 Census Mail List*. Washington, D.C.: General Accounting Office.

Total Error in the Dual System Estimator: The 1986 Census of Central Los Angeles County

MARY H. MULRY and BRUCE D. SPENCER¹

ABSTRACT

The U.S. Bureau of the Census uses dual system estimates (DSEs) for measuring census coverage error. The dual system estimate uses data from the original enumeration and a Post Enumeration Survey. In measuring the accuracy of the DSE, it is important to know that the DSE is subject to several components of nonsampling error, as well as sampling error. This paper gives models of the total error and the components of error in the dual system estimates. The models relate observed indicators of data quality, such as a matching error rate, to the first two moments of the components of error. The propagation of error in the DSE is studied and its bias and variance are assessed. The methodology is applied to the 1986 Census of Central Los Angeles County in the Census Bureau's Test of Adjustment Related Operations. The methodology also will be useful to assess error in the DSE for the 1990 census as well as other applications.

KEY WORDS: Nonsampling error; Post enumeration survey; Coverage evaluation, Undercount; Capture-Recapture.

1. INTRODUCTION

The dual system estimator (DSE) is used in several contexts for estimating the size of a population. Its applications range from wildlife populations to human populations. DSEs of births are used at the U.S. Bureau of the Census in the formation of the demographic analysis estimates of the national population. Currently, the Census Bureau intends to use DSEs for measuring coverage error in the 1990 Decennial Census. This paper focuses on the application of the DSE in the census context where the two systems are the original enumeration and a Post Enumeration Survey (PES).

The obvious estimator based on the DSE of census undercoverage is \widehat{UC} , given by $\widehat{UC} = \text{DSE} - \text{CEN}$, with CEN referring to the size of the original census enumeration. Since $\text{DSE} = \text{CEN} + \widehat{UC}$, the DSEs also provide alternative estimates of population. A more general class of alternative estimates based on the DSE (Spencer 1980; 1986) is $(1 - f) \times \text{CEN} + f \times \text{DSE}$, or equivalently

$$\text{CEN} + f \times \widehat{UC}$$

with $0 \leq f \leq 1$.

Estimates of total error of the DSE are essential for determining what value of f leads to the most accurate estimator of population size. Since the range of values for f include 0 and 1, the selection of either CEN or DSE is possible. The criteria for improvement of one set of population estimates over another may be based on measures of the quality of the distribution of the population (Hogan and Mulry 1987; Spencer 1986). Estimates of total error in the

¹ Mary Mulry, Undercount Research Staff, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C. 20233. Bruce Spencer, Department of Statistics, Northwestern University, Evanston, IL 60201 and NORC. The views expressed are attributed to the authors and do not necessarily reflect those of the Census Bureau.

DSE are also important for statistical planning purposes, *e.g.*, how much money should be spent and how big a sample should be fielded in the PES.

DSEs are subject to several components of nonsampling error, in addition to sampling error. We present models of the total error and the components of error in the DSE. The models relate observed indicators of data quality to the first two moments of the components of the error. We then use techniques of propagation of error to estimate the bias and variance of the DSE. In doing so, we assess the total error, or the joint effect of the errors. Previous work on error models for the DSE includes Seltzer and Adlakha (1974).

The methodology is applied to the 1986 Census of Central Los Angeles County, also known as the 1986 Test of Adjustment Related Operations (TARO) conducted in Los Angeles (Difendal 1988). The PES in TARO comprised about 6,000 housing units and over 19,000 people. A sensitivity analysis shows how the component errors interact, which ones cancel, and which ones compound each other. The methods described here to estimate the error in the TARO DSE can be extended to estimate the error in the 1990 DSEs.

We have tried to organize this paper to facilitate incomplete reading of the paper. Section 2 introduces and presents the rationale for the TARO DSE and its major components. Our strategy for assessing the component errors and combining them to estimate the total error in the DSE is described next (Section 3). A detailed description of the DSE, with notation, is necessary for precise description of the component errors (Section 4). Following that description is an assessment of the component errors (Section 5). A synthesis of the component errors leads to estimates of the total error of the DSE (Section 6). Our major conclusions are then presented (Section 7).

2. DUAL SYSTEM ESTIMATOR

The application of the dual system estimator requires assuming that there are two lists of the population. The first list is the original census enumeration, and the second is an implicit list of those covered by the sampling frame for the P sample of the PES, whom we will call the P-sample population. The sampling frame itself is not a list of people, but of census blocks.

The P sample is one of the two samples that comprise the PES. The PES is composed of the E sample, which is a sample of census enumerations, and the P sample, which is a sample of the population. The E sample is selected to estimate the number of enumerations that are erroneous. The P sample is selected to estimate, through dual system estimation, the number of people missed by the original enumeration.

Table 1
Probabilities of Inclusion in a Cell

	Original Enumeration		
	In	Out	Total
P sample In	p_{i11}	p_{i12}	p_{i1+}
Out	p_{i21}	p_{i22}	p_{i2+}
Total	p_{i+1}	p_{i+2}	p_{i++}

Table 2
True Population Size in Each Cell

	Original Enumeration		
	In	Out	Total
P sample In	N_{11}	N_{12}	N_{1+}
Out	N_{21}	(N_{22})	(N_{2+})
Total	N_{+1}	(N_{+2})	(N_{++})

The dual system estimator is based on a model that the probabilities that the i -th individual in the population of size N is in the census or not and in the P sample or not are as shown in Table 1 (Wolter 1986a); see Wolter (1986a) for discussion and references to earlier work. The true population size in each category is defined in Table 2.

In Table 2, $N_{++} = N$, the total population size. Even if we could observe the N_{ij} 's in the first row and first column, the N_{ij} 's in parentheses would not be observed directly, but would have to be estimated from the model. The DSE of N then would have the form $N_{1+}N_{+1}/N_{11}$, which we will refer to as the ideal DSE.

In estimating population size for measuring census coverage error, the N 's are replaced by estimates from the original enumeration and two sample surveys, the P sample and the E sample. The survey data are weighted by the reciprocals of the selection probabilities. In the following definitions, the estimates with “ $\hat{}$ ” reflect the possible presence of nonsampling error:

- N_p = the weighted number of P-sample selections
- \hat{N}_p = the estimate of the total population from the P sample.
- CEN = the size of the original enumeration
- II_1 = the number of persons imputed
- II_2 = the weighted number of census enumerations with insufficient information for matching
- EE = the weighted number of erroneous enumerations in the original enumeration, based on the E sample
- \widehat{EE} = the estimate of the number of erroneous enumerations in the original enumeration
- C = CEN - II_1 - II_2 - EE = the weighted number of distinct people in the original enumeration from the E sample,
- \hat{C} = CEN - II_1 - II_2 - \widehat{EE} = the estimate of the number of distinct people in the original enumeration from the E sample,
- M = the weighted number of people in the census and the P sample
- \hat{M} = the estimate of the number of people in the census and the P sample.

With this notation, \hat{N}_p estimates N_p , which unbiasedly estimates N_{1+} . The ratio \hat{C}/\hat{M} is used to estimate the ratio N_{+1}/N_{11} . (By themselves, \hat{C} and \hat{M} are not good estimators of N_{+1} and N_{11} .) Thus, the estimator has the form $\hat{N}_{++} = \hat{N}_p \hat{C}/\hat{M}$. The ratio \hat{C}/\hat{M} contains a correction for erroneous enumerations and for cases with insufficient information for matching, II_1 and II_2 , so that cases with no chance of being included in the denominator are also excluded from the numerator.

The DSE is used to estimate the percent net undercount, or the *net undercount rate*, in the original enumeration,

$$\hat{U} = 100 (\text{CEN} - \hat{N}_{++}) / \hat{N}_{++}.$$

For the TARO site (*i.e.* Central Los Angeles County) as a whole, $\text{CEN} = 355,352$, $\hat{N}_p = 336,707$, $\hat{C} = 343,567$, $\hat{M} = 298,204$, and $\hat{N}_{++} = 388,040$. Using these numbers, the estimate of the net undercount rate is 8.42.

3. STRATEGY FOR ASSESSING TOTAL ERROR

The DSE is subject to various sources of error, including error due to incorrect addresses from the P sample, error due to missing data (unit and item nonresponse), response errors, interviewer errors, correlation bias, sampling error, *etc.* We wish to estimate the effects of these diverse sources of error on the DSE.

The first step in our strategy is to express the DSE as a function of components. We have constructed the components so that, for the most part, the different sources of error act either independently or perfectly dependently on different components. By isolating the effects of the various errors, we are better able to identify the major distinct sources of error.

Next, we estimate the first two moments of the component errors, one component at a time. In doing so we draw upon the results of various TARO evaluations and quality control programs. The way we constructed the components implies that correlation between component errors typically equals either 0 or 1.

To study the propagation of errors we have used computer simulation methods. A multivariate distribution of the error components, say F , was assumed. The specification of F was consistent with the first two moments as estimated in Section 5. Realizations of the component errors were simulated by pseudo-random draws from F and then the DSE was calculated; this procedure was repeated 10,000 times and the resulting empirical distribution of the DSE was used as an estimate of its actual distribution. The first two moments of the latter distribution provide numerical estimates of the total error of the DSE.

Sensitivity analysis was performed to discover the importance of using one distributional form for F rather than another. The results suggest that the exact distributional form (beyond the first two moments) is relatively unimportant (see Section 6).

We adopted a Bayesian approach in investigating of the error in the DSE. We estimated the first two moments of the distributions for the error components, then we derived the posterior distribution of the undercount rate conditional on the observed values of \hat{C} , \hat{N}_p , \hat{M} , *etc.*

4. COMPONENTS OF THE DSE

The DSE is subject to sampling errors and nonsampling errors, including failure of assumptions underlying the DSE model. The DSE does have a bias, but the bias in the census context is negligible (Wolter 1986a). Nonsampling errors may affect the accuracy of estimation of N_{+1} , N_{1+} , and N_{11} . Descriptions of the nonsampling error follow.

The error in the estimation of N_{+1} is defined by $\hat{C} - N_{+1} = (\hat{C} - C) + (C - N_{+1})$. The first term $(\hat{C} - C)$ is the net nonsampling error, which contributes to both bias and variance, and the second term $(C - N_{+1})$ is the sampling error, which contributes only to the variance. Define the net nonsampling error as $c = \hat{C} - C$.

The net error c arises during the processing of the E sample when respondents are misclassified as to whether they are correctly or erroneously enumerated in the original enumeration. Therefore, c has three components: c_e , which occurs during the data collection and processing; c_b , caused by a PES design that fails to balance estimates of the gross overcount and gross undercount; and c_i , caused by missing data, $c = c_e + c_b + c_i$. Sections 5.5, 5.6 and 5.7 cover c_e , c_b and c_i , respectively.

The error in the estimation of N_{1+} is defined by $\hat{N}_p - N_{1+} = (\hat{N}_p - N_p) + (N_p - N_{1+})$. The first term $(\hat{N}_p - N_p)$ is the nonsampling error, which contributes to both bias and variance and the second term $(N_p - N_{11})$ is the sampling error, which contributes only to the variance. The net nonsampling error is defined by $n_p = \hat{N}_p - N_p$.

The net error n_p arises during the interviewing for the P sample when the P-sample selections are not interviewed. This situation occurs when household members are fabricated or when there is missing data. Therefore, n_p has two components: n_{pf} , the error due to fabrication and n_{pi} , the error due to missing data, $n_p = n_{pf} + n_{pi}$. Section 5.3 discusses n_{pf} , and Section 5.7 covers n_{pi} .

The error in the estimation of N_{11} is defined by $\hat{M} - N_{11} = (\hat{M} - M) + (M - N_{11})$. The first term $(\hat{M} - M)$ is the net nonsampling error, which contributes to both bias and variance, and the second term $(M - N_{11})$ is the sampling error, which contributes only to the variance.

To facilitate the description of the nonsampling error in the estimation of N_{11} , consider the following tables of P-sample selections and respondents. Entries in Table 3 are the weighted number of P-sample selections in each category. Entries in Table 4 are the weighted number of P-sample responses in each category. Entries in Table 5 are estimates of the number of people in each category based on the P-sample interviewing, responses, and matching operation.

Table 3
P-sample Selections

P-sample Selections	Census Enumeration Status	
	Enumerated	Not Enumerated
Not reported	D_{11}	D_{12}
Reported		
Correct Census Day Address	D_{21}	D_{22}
Wrong Census Day Address	D_{31}	D_{32}

Table 4
Enumeration Status of P-sample Respondents

P-sample Status	Census Enumeration Status	
	Enumerated	Not Eumerated
Fabricated	A_{11}	A_{12}
Not Fabricated		
Correct Census Day Address	A_{21}	A_{22}
Wrong Census Day Address	A_{31}	A_{32}

Table 5
Match Status of P-sample Respondents

P-sample Status	Match Status	
	Matched	Not Matched
Fabricated	B_{11}	B_{12}
Not Fabricated		
Correct Census Day Address	B_{21}	B_{22}
Wrong Census Day Address	B_{31}	B_{32}

Since the P-sample selections who appear as reported in Table 3 are the respondents who are not fabricated in Table 4, $D_{21} = A_{21}$ and $D_{31} = A_{31}$. Also, $A_{11} = 0$ since a case fabricated during the PES cannot be enumerated in the census. Therefore,

$$M = D_{11} + D_{21} + D_{31} = D_{11} + A_{21} + A_{31}.$$

Since a case fabricated during the PES would not have a corresponding census enumeration, we assume $B_{11} = 0$. Therefore, $\hat{M} = B_{11} + B_{21} + B_{31} = B_{21} + B_{31}$.

Then the nonsampling error in the estimation of N_{11} , called m , may be defined as follows:

$$\begin{aligned} m &= \hat{M} - M \\ &= (B_{11} + B_{21} + B_{31}) - (D_{11} + D_{21} + D_{31}) \\ &= -D_{11} + (B_{21} - A_{21}) + (B_{31} - A_{31}). \end{aligned}$$

The error m has three components: $(B_{21} - A_{21})$, which is the error introduced in the matching operation (Section 5.2); $(B_{31} - A_{31})$, which is the error introduced by respondents giving the wrong Census Day address (Section 5.3); and $-D_{11}$. D_{11} has two components: missing match status m_i and fabrication m_f . Section 5.7 covers missing match status, and Section 5.4 covers fabrication.

The ideal DSE can be written as follows:

$$N_{1+} N_{+1}/N_{11} = (\hat{C} - c)(\hat{N}_p - n_p)/(\hat{M} - m).$$

5. COMPONENTS OF PES ERROR

Estimates of the first two moments of the posterior distribution of the undercount rate derive from estimates of the first two moments of the components of PES error. The components are correlation bias, matching error, accuracy of the reported Census Day address, fabrication in the P sample, measurement of erroneous enumerations, balancing the estimates of the gross overcount and the gross undercount, missing data, and sampling error. We next describe the source of each component of PES error and give models for each component. We model the component errors in terms of observable indicators of data quality. We estimate the first two moments of the distributions of the errors for use in the total error model in Section 6.

5.1 Correlation Bias

5.1.1 Source of Error

An important concern for dual system estimation is that the estimate of the proportion of the population enumerated in the census, based on the P sample, is accurate. The violation of one of the independence assumptions underlying dual system estimation may cause the estimate of the proportion of the population enumerated in the census, and thereby the estimate of the population, to be biased.

Three independence assumptions are made for dual system estimator:

Causality. The event of being included in the census is independent of the event of being included in the PES. That is, the cross-product ratio satisfies

$$\theta_i = p_{i11} p_{i22} / p_{i12} p_{i21} = 1, \text{ for } i = 1, \dots, N.$$

Homogeneity. The capture probabilities satisfy $p_{i1+} = p_{1+}$ or $p_{i+1} = p_{+1}$ for $i = 1, \dots, N$, within each of the post-strata.

Autonomy. The census and the PES are created as a result of N mutually independent trials.

The homogeneity assumption follows combination model M_{th} in Wolter (1986a). All the development for the Peterson model M_t in Wolter (1986a) also applies to model M_{th} when enough information is available to form post-strata where M_t holds.

To control heterogeneity in the population the Census Bureau post-stratifies the data based on demographic and geographic variables, a technique originally recommended by Sekar and Deming (1949). An estimate of the population in each post-stratum is calculated and then all the estimates are summed to give an estimate of the total population. Unless the failure of the homogeneity assumption is severe, the estimate lies between the census and the truth.

Research by Wolter (1986b) and Cowan and Malec (1986) has demonstrated that the failure of the autonomy assumption has a negligible effect on the bias of the DSE but causes an increase in its variance. Wolter's formulation allows household members to act individually (autonomy) or together (failure of autonomy). Cowan and Malec present a model that permits clustering of the census misses (failure of autonomy). Next, we model the combined effect of the sources of correlation bias on the DSE.

5.1.2 Definition

For insight into the effect of correlation bias, assume all $\theta_i = \theta$ and write the true population size as

$$N = N_{11} + N_{12} + N_{21} + \theta (N_{12}N_{21}/N_{11}),$$

where θ_i is the cross-product ratio defined in Section 5.1.1.

The correlation bias affects only the last term because the other three may be estimated directly. The parameter θ represents the effect of the failure of the independence assumptions. When the independence assumptions hold, $\theta = 1$.

The correlation bias, arising when θ does not equal 1, is the only contributor to t , the error due to failure of the model. The population size can be written as follows:

$$\begin{aligned} N &= N_{1+}N_{+1}/N_{11} + t \\ &= N_{1+}N_{+1}/N_{11} + (\theta - 1)(N_{12}N_{21}/N_{11}). \end{aligned}$$

Therefore, the correlation bias, $t = (\theta - 1)(N_{12}N_{21}/N_{11})$.

5.1.3 Measurement

The parameter θ may be estimated at the national level for racial and ethnic subgroups using demographic analysis estimates of the population size. Note, however, that this technique presumes that the demographic analysis estimates are accurate. Even so, this formulation also permits varying θ to assess the sensitivity of the DSE to the estimate of the effect of the violation of the independence assumptions.

5.1.4 Estimation

Estimates for θ were not made for the 1986 TARO because an alternate source for population estimates did not exist, *e.g.*, no demographic analysis estimates were feasible. However, Ericksen and Kadane (1985) made three estimates of θ for blacks for the 1980 census: 2.1, 2.7, and 3.7. Since the population in the 1986 TARO was predominantly minority (73 percent Hispanic, 12 percent Asian, and 15 percent non-Asian and non-Hispanic), the Ericksen and Kadane estimates for 1980 will be used in this paper: $E(\theta) = 2.1, 2.7, \text{ or } 3.7$, $\text{Var}(\theta) = 0$. We are treating θ as fixed, but unknown. A sensitivity analysis is conducted in Section 6 to demonstrate the effect of alternative values of θ .

These estimates of θ are consistent with the reports of the participant observers in the Los Angeles test site (Childers *et al.* 1987). Our professional judgment is that correlation bias is higher for urban areas than for the country as a whole. This implies that these estimates may be conservative for the Los Angeles test site because it was urban.

5.1.5 Summary

In the total error model the first two moments of the posterior distribution of the correction factor for correlation bias are assumed to be $E(\theta) = 2.1, 2.7, \text{ or } 3.7$, and $\text{Var}(\theta) = 0$.

5.2 Matching Error

5.2.1 Source of Error

Matching error in this discussion refers to errors that occur in the operation where the P sample is matched to the original enumeration. Therefore, matching error does not encompass response errors that arise in the data collection. Although other types of errors may result in an inaccurate assignment of a P-sample respondent's census enumeration status, these sources are treated in other components of error.

After the P-sample interviewing is completed, a search of the census is conducted to determine if the respondents are enumerated. Then the P-sample respondents are designated as matching an enumeration in the census or as not enumerated in the census. Errors in assigning the enumeration status to P-sample persons which occur during the processing of the data are known as matching error. Errors may occur in either direction. People may be designated as matching a census enumeration although they are not in the census, called a "false match," or people may be designated as not enumerated although they are, called a "false nonmatch." Matching error will cause a bias in the estimate of the number of people in both the census and the P-sample population and thereby introduce a bias into the estimates of the number of people missed by the census.

5.2.2 Definition

The denominator N_{11} of the dual system estimator is estimated from sample survey data, the P sample. The following were introduced in Section 4:

A_{21} = the weighted number of people who were enumerated,

B_{21} = the estimate of the number of people who match.

Then the net error due to incorrect classification of enumeration statuses, m_m , may be defined as $m_m = B_{21} - A_{21}$. The conditional expected value and variance of m_m given observed value \hat{M} are denoted by $E(m_m)$ and $\text{Var}(m_m)$.

5.2.3 Measurement

Measurement of m_m is possible by processing a sample of the cases a second time *i.e.*, by having highly trained personnel rematch them. The assumption underlying an independent rematch of a sample is that the personnel with more training make fewer mistakes in classifying enumeration statuses although they have the same materials and information available as the original workers. The original match codes and the evaluation match codes can be reconciled, and the discrepancies can be resolved.

Two evaluations of the clerical matching were conducted with the 1986 TARO data. One study evaluated the clerical matching for movers, and another evaluated the clerical matching for nonmovers.

In the evaluation of matching for nonmovers (Corby and Mulry 1988), a probability subsample of 35 blocks was chosen for a rematch by professionals from headquarters. The sample was stratified by match rate, and blocks with low match rates were sampled at a disproportionately high rates so that the quality control staff could learn as much as possible about matching errors. Adjacent blocks were not searched so the false nonmatches are possibly underestimated.

The second evaluation study considered matching error for movers (Childers *et al.* 1987). There were 90 movers who were not matched in TARO, and all of these movers were rematched. Eleven matches were found, two of which had been lost during the computer editing.

5.2.4 Estimation

We now use the results of the evaluation subsamples to estimate the moments of the distribution of m_m from the PES sample. Not conducting an extended search in the evaluation for the nonmovers probably reduced the number of false nonmatches found. Experience with extended searches implies that adding an additional 20 percent of the net error of 70 (Hogan and Wolter 1988) is a conservative way to compensate for the lack of one. The results from the two evaluations yield a net error of 95 in the PES sample. Therefore, the net error rate is $-.0055$. We apply the net error rate to only the P-sample cases with a resolved match status because the error in the imputation for the unresolved cases is covered in the Missing Data Section 5.7. The expected value of m_m becomes $E(m_m) = -1831$, when the overall sampling weight of 17 is used.

An estimate of the variance of the estimate of net matching error for nonmovers has not been calculated. The sample variance of the number of errors for movers is zero because all the nonmatched movers were rematched. However we do not believe that the true variance is zero. One way to obtain a variance specification would be to assume that the errors occurred in the manner of a mixture of Poisson processes, *e.g.*, matching errors for movers followed one Poisson process and matching errors for nonmovers independently followed another Poisson process. Treating the errors as arising from a simple Poisson process would then lead to a conservative estimate of variance; in this case the variance would be estimated by 17×107 . However, the Poisson model may not be conservative if the errors occur in clusters. In an attempt to develop conservative estimates of variance, we have (somewhat arbitrarily) multiplied the variance estimate under the simple Poisson model by the overall sampling weight to obtain

$$\text{Var}(m_m) = (17)^2 \times 107 = 30,923.$$

5.2.5 Summary

For the total error model, the first two moments of the posterior distribution of the net matching error for the PES sample are assumed to be $E(m_m) = -1831$ and $\text{Var}(m_m) = 30,923$.

5.3 Quality of the Reported Census Day Address

5.3.1 Source of Error

Some of the respondents in the P sample have moved between Census Day and their PES interview. The respondents may misreport whether they have moved during the time lapse. If they have moved, they may not report their previous address accurately, or their previous address may not be geocoded correctly by the staff. Any of these types of errors may cause the matching operation to search the census in an area other than where the respondent was enumerated. These errors may lead to assigning a nonmatch status to respondents who actually were enumerated because the matching operation is unable to locate their enumerations. Inappropriate assignment of the status of nonmatch will cause the estimate of the number of people missed by the census to be biased upward.

Circumstances under which inaccurate reporting of the Census Day address by a PES respondent will not cause a false nonmatch do exist. If the Census Day address is inside the search area for the reported address, and the reported address is geocoded correctly, then the matching operation will find the person.

5.3.2 Definition

The denominator N_{11} of the dual system estimator is estimated from sample survey data, the P sample. The following were introduced in Section 4:

A_{31} = the weighted number of people with an inaccurate Census Day address who are enumerated,

B_{31} = the estimate of the number of people with an inaccurate Census Day address who match at another address.

Then the net error due to inaccurate reporting of the Census Day address, m_a , may be defined as $m_a = B_{31} - A_{31}$. The conditional expected value and variance of m_a given the observed value \hat{M} are denoted by $E(m_a)$ and $\text{Var}(m_a)$.

5.3.3 Measurement

Measurement of m_a is based on a follow-up of a sample of P-sample respondents whose enumeration status is "not enumerated". Data from the follow-up are used to estimate the error that arises when people who were enumerated misreport their Census Day address when they respond to the PES.

An evaluation of the quality of the reporting of the Census Day address was conducted after the 1986 TARO. A post-production follow-up which reinterviewed a sample of 903 of the non-matches was aimed at determining the number of nonmatches caused by misreporting mover status. Another search to match respondents who reported they in fact had moved within the test site was made at the new address.

5.3.4 Estimation

The sample cases found to have errors in their reported Census Day address may be used to estimate

L_e = the weighted number of people who erroneously report their Census Day address in their P-sample interview.

A search of census enumerations at the newly reported addresses produces

r_{am} = the estimator of the percentage of people with errors in the location of their reported Census Day address who match census enumerations.

Then the expected value of the error m_a is estimated by

$$E(m_a) = - r_{am}L_e.$$

The results of the post-production follow-up (Hogan and Wolter 1988) yielded a misreporting rate of at most 3.1 percent in the P sample. A match rate of 33 percent was estimated for those who misreported their Census Day address and moved within the test site. If we assume the match rate for those who reported a census day address outside the test site is also 33 percent, then the expected value $E(m_a) = -3481$.

An estimate of the variance of the error due to misreporting has not been made. Our professional judgment is that a conservative estimate of the variance at the PES sample level is 900. Therefore, the variance at the TARO site level is

$$\text{Var}(m_a) = (17)^2 \times 900 = 260,100.$$

5.3.5 Summary

For the total error model, the first two moments of the distribution of the error due to misreporting of Census Day address for the PES sample are assumed to be $E(m_a) = -3481$ and $\text{Var}(m_a) = 260,100$.

5.4 Fabrication in the P sample

5.4.1 Source of Error

Interviewers may fabricate people in P-sample housing units. Research has shown that interviewer fabrication during the PES may result in a substantial bias in the estimates of census coverage error based on the dual system estimator. Basically, the creation of fictitious individuals may decrease the PES match rate, causing the estimate of coverage error to be too large.

Experience at the Bureau of the Census has shown that fabrication of the members of a whole household is the problem for household surveys. Rarely is there a fabrication of the household member in a household where the other members are the real residents.

The quality control operation for the interviewing phase of the P sample is designed to check for fabricated interviews and to interview the real household members. Therefore, no statistical correction for fabrication in the P sample is made in the formation of the dual system estimates.

5.4.2 Definition

The N_{11} and N_{1+} in the dual system estimator are estimated from sample survey data, the P sample. The following were introduced in Section 4:

m_f = the weighted number of people who were replaced by fabricated P-sample interviews and who were enumerated,

n_{pf} = the error in N_{pf} due to households that were fabricated in the P sample.

The posterior expected values and variances of m_f and n_{pf} are denoted by $E(m_f)$ and $E(n_{pf})$ and $\text{Var}(m_f)$ and $\text{Var}(n_{pf})$.

5.4.3 Measurement

In the 1986 TARO, the estimate of the fabrication rate based on the quality control of the interviewing was approximately 0.6 percent. The estimate of the fabrication rate based on a post-production follow-up was approximately 1.2 percent (Hogan and Wolter 1988).

5.4.4 Estimation

We now estimate the moments of the posterior distributions of n_{pf} and m_f from the PES sample. We believe it is reasonable to assume n_{pf} is negligible in TARO. Therefore, the expected value and variance are given by $E(n_{pf}) = 0$ and $Var(n_{pf}) = 0$.

The quality control data may be used to estimate r_f = the rate at which P-sample interviews are fabricated.

The search of the census enumerations for people in the P sample who were found by the quality control operation to not have been properly interviewed produces r_{fm} = the match rate for people not interviewed because their household was fabricated in the P sample.

In TARO, records were not kept so that the people who were discovered by the quality control not to have been interviewed properly could be identified. Therefore, no search was made for matching enumerations. Since we have no data available for a direct estimate of r_{fm} , we conservatively assume that the people not interviewed properly are like the people who were. We set r_{fm} equal to the final overall P-sample match rate.

We use the conservative results from the post-production follow-up to yield a fabrication rate of 1.2 percent. The match rate for TARO is 88.6 percent (Diffendal 1988). Therefore, the expected value of the error m_f is given by $E(m_f) = -2502$.

An estimate of the variance of the estimate of fabrication error has not been calculated. Our professional judgment is that a conservative estimate of the variance can be derived by the reasoning discussed in Section 5.4.2. Thus, we estimate that the variance for the TARO site is

$$Var(m_f) = (17)^2 \times 206 = 59,534.$$

5.4.5 Summary

For the total error model, the first two moments of the distribution of the net error due to fabricated interviews are assumed to be $E(m_f) = -2502$ and $Var(m_f) = 59,534$. The net error due to fabricated interviews in is assumed to be negligible, and therefore, $E(n_{pf}) = 0$ and $Var(n_{pf}) = 0$.

5.5 Measurement of Erroneous Enumerations

5.5.1 Source of Error

Some enumerations may have been entered in the census as the result of mistakes. These enumerations are called erroneous enumerations. Since the dual system estimator requires estimating the number of distinct people captured in the census, a correction is made for erroneous enumerations in the estimate of total population. Subtracting the estimate of the number of enumerations that do not correspond to distinct people from the census count provides an improved estimate of the number of distinct people captured in the census. This estimated correction is obtained from the E sample in the PES.

The following types of enumerations are considered erroneous: (1) people who died before Census Day, (2) people who were born after Census Day, (3) enumerations that do not refer to real people, (4) people duplicated, (5) people enumerated outside the search area where the matching operation looks for their enumeration. The search area for a case includes the block for its address and the ring of adjacent blocks.

This component is caused by errors in measuring census error. An error in the estimation of the number of erroneous enumerations occurs either when an enumeration in the E sample

is designated as erroneous although it is correct, or when an enumeration is designated as correct although it is really erroneous. Therefore, both positive and negative error can occur in the estimation of the number of erroneous enumerations.

The types of enumerations that are the most vulnerable to misclassification as to whether they are erroneous include the duplicated and fabricated enumerations. These errors are the only ones considered because the others are either inconsequential or are treated separately. Errors in identifying enumerations for people who died before Census Day and people who were born after Census Day have a trivial effect. Errors in classifying the enumeration status because a person was enumerated outside the search area is covered in Section 5.6 on balancing the estimates of the gross overcount and the gross undercount.

5.5.2 Definition

The bias in the DSE due to misclassification of enumeration status is caused by error in the estimation of N_{+1} . In the formation of the estimate of the number of distinct people in the original enumeration \hat{C} , a correction is made for the number of erroneous enumerations, \overline{EE} . \overline{EE} and therefore \hat{C} are estimated from sample survey data, the E sample. Errors in the estimate \hat{C} occur through the misclassification of the enumeration status of E-sample cases. Let

c_e = the difference between the weighted number of erroneous enumerations misclassified as correct and the weighted number of correct enumerations misclassified as erroneous.

The expected value of c_e , conditional on the observed value \hat{C} , is denoted by $E(c_e)$. The variance of c_e , conditional on the observed value \hat{C} , is denoted by $\text{Var}(c_e)$.

5.5.3 Measurement

Processing error may be measured directly using a rematch of a sample of cases. Errors from other sources, such as duplications due to violations of census residency rules, can be assessed by viewing the frequency distributions of the erroneous enumerations. This is preferable to direct measurement of these errors because of the difficulties in obtaining accurate data in additional follow-ups. When tests confirm that the gross errors from these sources are under control, the net error can be assumed to be negligible. For example, the distribution of the erroneous enumerations by age group is expected to have a large number of duplications in the highly-mobile groups of the population where there are more opportunities for the census residency rules not to be followed.

In the 1986 TARO, an evaluation of the E-sample processing was conducted in conjunction with the evaluation of the P-sample matching operation discussed in Section 5.2.3 (Corby and Mulry 1988). The data for the E sample from the same subsample of 35 blocks were reprocessed.

5.5.4 Estimation

We now estimate the moments of the distribution of c_e from the PES sample. The results of the reprocessing (Hogan and Wolter 1988) yield a net error rate of 0.0007 in the identification of correct enumerations. The expected value of c_e is $E(c_e) = -238$. This estimate is based on the E sample with a resolved enumeration status because the error in the imputation for the unresolved cases is covered in the Missing Data Section 5.7.

An estimate of the variance of net error has not been calculated. Our professional judgment is that a conservative estimate of the variance can be derived by the reasoning discussed in Section 5.2.2. Thus, we estimate that the variance for the TARO site is $\text{Var}(c_e) = (17)^2 \times 14 = 4,046$.

5.5.5 Summary

For the total error model, the first two moments of the posterior distribution of the net error in identifying correct enumerations are assumed to be $E(c_e) = -238$ and $\text{Var}(c_e) = 4,046$.

5.6 Balancing the Estimates of the Gross Overcount and Undercount

5.6.1 Source of Error

Both the E sample and the P sample measure enumeration errors in the census. The E sample measures the gross overcount in the form of erroneous enumerations. The P sample measures the gross undercount in the form of those not enumerated. Ideally, the entire census would be searched before a P-sample person was declared to be not enumerated. Ideally, the entire country would be searched to determine if an E-sample enumeration is a duplicate or fictitious. Of course, such extensive searches are simply not feasible in the performance of the PES. These searches must be limited in the reasonable manner. The way chosen has to preserve the net error although the measured gross overcount and the measured gross undercount may increase due to limiting the search area. The gross overcount and the gross undercount have to balance to equal the net coverage error.

Failure to have procedures which balance the estimated gross overcount and the estimated gross undercount may cause an incorrect number of enumerations in the E sample to be designated as erroneous when they are correct. This error may cause either an upward or downward bias.

Balancing is not an issue for the design of the PES planned for 1990 and tested in the 1986 TARO, as it was in 1980. The design calls for overlapping the P sample and the E sample. The same blocks are included in the P sample as in the E sample. The P-sample search area is, by definition, the proper search area. The E-sample search area is chosen to be consistent with the P-sample search area.

5.6.2 Summary

Error due to geocoding error is believed to be negligible in the 1986 TARO and will not be included in the total error model. The appendix contains a model for balancing error.

5.7 Missing Data

5.7.1 Source of Error

Both the E sample and the P sample have missing data. The E sample has cases where the information required to determine whether the person is correctly or erroneously enumerated in the census is not available. The P sample has cases where the information needed to determine whether the person is enumerated in the census is not available. The probability of being enumerated is imputed statistically to compensate for the inability to resolve the case.

An unresolved status may occur in more than one way. The interviewer may be unable to obtain an interview during the P-sample interviewing or during the PES follow-up. A P-sample or E-sample questionnaire may not have all the demographic and housing information required for the estimation. Even with all the information requested on the questionnaires, the circumstances may be so unclear that the enumeration status can not be resolved.

5.7.2 Measurement

We assess the error in the DSE caused by missing data instead of considering each component c_i , m_i and n_{pi} separately. Our approach is to perform a sensitivity analysis of reasonable alternative models for compensating for missing data. First a preferred method of imputation for

unresolved P-sample and E-sample enumeration statuses is specified prior to the implementation of the PES. Reasonable alternative treatments of the missing data can be suggested by problems that arise during the collection and processing of the PES data. The DSE can be computed under these alternative models for compensating for missing data. The range of the alternative estimates indicates the sensitivity of the DSE to the method of imputation. For example, a narrow range implies that the estimates are robust, and the missing data cause little uncertainty in the estimates.

5.7.3 Estimation

The effect of missing data on the estimates from the 1986 TARO was assessed by examining the range of estimates obtained when methods of imputation based on reasonable alternative assumptions were used in place of the preferred method. These included alternative treatment of proxy responses, movers, and designation of fictitious enumerations (Schenker 1988). The alternative treatment of the proxy interviews for P-sample cases classified them as noninterviews and applied the weighting adjustment. This essentially assigned proxy cases the same match rate as nonproxy cases. The alternative treatment of the P-sample movers reclassified them all as unresolved and imputed a match probability, instead of imputing for only those who were not resolved. This essentially assigned movers the same match rate as nonmovers. The alternative treatment of fictitious cases resulted from a review of the unresolved E-sample cases by experienced matching personnel who converted some unresolved cases to fictitious. This raised both the observed and imputed rates of erroneous enumeration.

Models 000 and 111 shown in Table 4 of Schenker's paper give the upper and lower bounds of the estimates of undercount rates, respectively. Both models differ from TARO in that they have in-movers as substitutes for out-movers. P-sample in-movers are P-sample respondents who moved into their housing unit between Census Day and PES interviewing. In the 1986 TARO the P-sample in-movers from areas outside the test site were omitted from the PES estimation. The omission of the out-movers from estimation essentially assumes that they had the same capture rate in the original enumeration as the included cases. Movers are believed to have a lower capture rate than nonmovers. Model 000 has the TARO treatments while Model 111 has all the alternative treatments.

5.7.4 Summary

The effect of missing data on the distribution of the total error is assessed by computing the distribution of the undercount rate under several reasonable imputation methods. The alternative methods which yield the upper and lower bounds for the undercount are used in the total error analysis.

5.8 Sampling Error

5.8.1 Source of Error

The observed DSE is subject to sampling error because \hat{N}_p , \hat{C} , and \hat{M} are estimated from samples. The sample size for the PES is determined by the amount of sampling error and budget allowable. Other things being equal, the larger the sample size the lower the amount of sampling error introduced in the estimates. The sampling error is affected by the estimator and the sampling design. In the TARO PES design, both the P-sample and the E-sample observations are collected from the same sample of blocks. All the people residing in the housing units in the selected blocks are included in the P sample. All enumerations assigned by the census process to the sample block are included in the E sample. The estimation of the sampling error takes into account the tendency for census misses and erroneous enumerations to be correlated within blocks and within housing units. Experience has shown that many hard-to-enumerate areas have both a higher rate of omissions and a higher rate of erroneous enumerations.

5.8.2 Measurement

The standard randomization theory model for survey sampling is appropriate for estimating the variance of the DSE. The coefficient of variation which is the ratio of the square root of the variance of the observed DSE to the mean of the distribution of the DSE provides information on the amount of sampling error in the DSE.

The Taylor series estimator of variance for the observed dual system estimator (Moriarity 1987), $v(\hat{N}_{++})$, is given by

$$\begin{aligned} v(\hat{N}_{++}) &= \hat{N}_{++}^2 (v(\hat{N}_p)/\hat{N}_p^2 + v(\hat{M})/\hat{M}^2 - 2c(\hat{N}_p, \hat{M})/\hat{N}_p\hat{M}) \\ &\quad + \hat{N}_p^2 v(\hat{E})/\hat{M}^2 + 2\hat{N}_{++} (\hat{N}_p c(\hat{E}, \hat{M})/\hat{M}^2 - c(\hat{E}, \hat{N}_p)/\hat{M}), \end{aligned}$$

where

$$\begin{aligned} \hat{E} &= II_2 + \widehat{EE}, \\ v(X) &= \text{the estimator of the variance of an estimator } X, \\ c(X, Y) &= \text{the estimator of the covariance between } X \text{ and } Y. \end{aligned}$$

The categories II_2 , insufficient information for matching, and \widehat{EE} , erroneous enumerations, are treated as one group in the variance estimation. The variance and covariance estimators reflect the cluster sampling of blocks and block clusters.

5.8.3 Estimation

The standard deviation of the dual system estimate of 388,040 for the TARO site is 3,100.37. The coefficient of variation is 0.008. This implies the standard deviation for the estimated net undercount rate is 0.7 percent.

5.8.4 Summary

The sampling error for the TARO DSE is 3,100.37, and the sampling error for the TARO net undercount rate estimate is 0.70 percent.

6. SYNTHESIS OF TOTAL ERROR

The combined effect of the component errors will be summarized by posterior distributions for the net undercount rate. The bias in the estimate of net undercount rate, $B(U)$, is estimated by the difference between and the mean of the posterior distribution. To construct the posterior distribution, we used a simulation method with 10,000 repetitions, generating pseudo-random component errors and adding them to the TARO estimates. Using the formulas in Section 5.1.2, we obtain the following formula:

$$\begin{aligned} N &= (\hat{N}_p - n_p) + (\hat{C} + c - (\hat{M} - m)) \\ &\quad + \theta(\hat{C} - c - (\hat{M} - m))(\hat{N}_p - n_p - (\hat{M} - m))/(\hat{M} - m) \\ &= (\hat{C} - c)(\hat{N}_p - n_p)/(\hat{M} - m) \\ &\quad + (\theta - 1)(\hat{C} - c - (\hat{M} - m))(\hat{N}_p - n_p - (\hat{M} - m))/(\hat{M} - m). \end{aligned}$$

Several different distributions were used to reflect alternative estimates of imputation error, alternative estimates of correlation bias (parameterized by θ), and alternative marginal distributional forms for the components – normal, gamma, and uniform.

In this study, the estimate of percent net undercount for the TARO site is 8.42 with a sampling standard deviation of 0.7. This estimate was selected because estimates of nonsampling error components are available only for the site as a whole. When a DSE is constructed for each post-stratum and then the DSEs are summed to give an estimate for the site, the percent net undercount estimate is 9.02.

Table 6 displays the means and standards deviations of the error components for the PES sample. Recall that the DSE for the TARO site is 388,040, $\hat{M} = 298,204$, $\hat{C} = 343,567$, and $\hat{N}_p = 336,707$. The overall sampling weight, 17, was used consistently throughout all the simulations so that comparisons of the effect of alternative assumptions such as correlation bias parameter values, error distributions, and imputation models are appropriate. The methodology generalizes to other applications where a different sampling weight is used in each stratum.

Table 7 displays the effects of the individual errors on the posterior distribution of the undercount when the TARO imputation is used. The net matching Census Day address, and fabrication errors are all errors in \hat{M} . Therefore, the presence of only one of them alone causes the bias in the estimate of percent net undercount to be positive. The net E-sample error is an error in \hat{C} . The presence of E-sample error alone causes the bias in the estimate of percent net undercount to be negative. The estimate for correlation bias, was chosen to be 2.7, the median of Ericksen and Kadane's estimates. The presence of only correlation bias causes the bias in the percent net undercount estimate to be negative.

Table 6
Assumed Distributions of Error Estimates

	Mean	Standard Deviation
Net Matching Error	-1831	176
Census Address Error	-3481	510
Fabrication Error	-2502	244
Net E sample Error	-238	64

Table 7
Individual Effects of Errors on Posterior Distribution
of Percent Net Undercount and Bias
in the Estimate of Undercount

	E(U)	Std. Dev.	B(U)
Net Matching	7.86	0.06	0.56
Census Address	7.35	0.16	1.07
Fabrication	7.34	0.08	1.08
Net E sample	8.49	0.02	-0.07
Correlation Bias (2.7)	10.61	0.00	-2.19

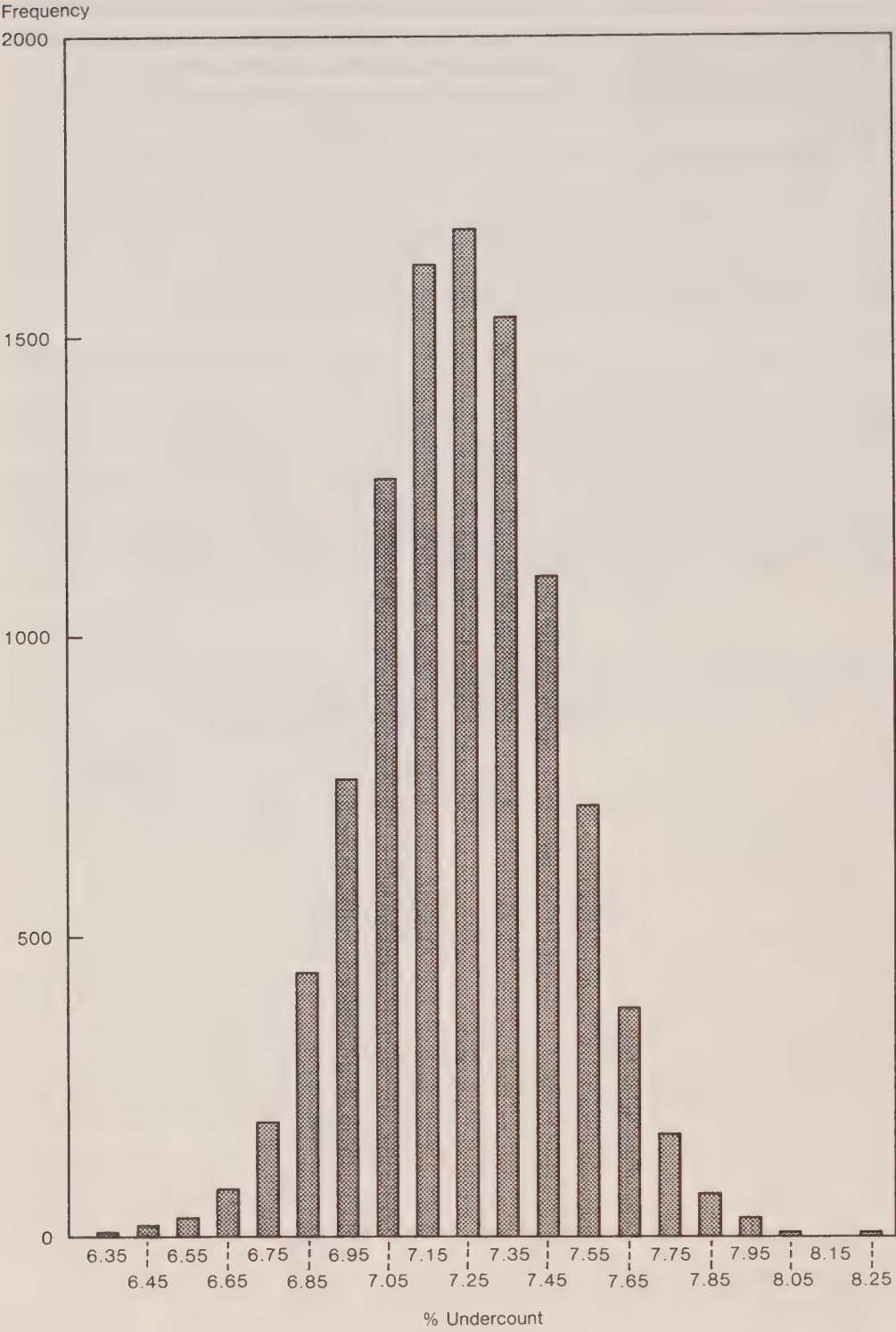


Figure 1. Percent Undercount when $\theta = 2.7$

Table 8
Percentiles of the Posterior Distribution of
Percent Net Undercount for $\theta = 2.7$

	1	5	10	25	50	75	90	95	99
Normal	6.70	6.86	6.94	7.08	7.24	7.40	7.54	7.63	7.79
Uniform	6.75	6.86	6.93	7.07	7.24	7.42	7.55	7.62	7.73
Gamma	6.67	6.84	6.93	7.08	7.24	7.40	7.53	7.61	7.74

Table 9
Posterior Distribution of the Net
Undercount Rate for Several Values of θ

θ	$E(U)$	St. Dev.	$B(\hat{U})$
1.0	5.75	0.18	2.67
2.1	6.72	0.22	1.70
2.7	7.24	0.23	1.18
3.7	8.09	0.27	0.33

Simulations were conducted where the first two moments for error n_p, c_e, m_m, m_f, m_a , and θ were held constant, but the distributions were varied. We assessed the total error when all the error distributions were normal, all were gamma, and all were uniform. Varying the distributions had minor effects on the distribution of the percent net undercount. In each case the distribution of the percent net undercount was very close to normal. Figure 1 shows the distribution of the undercount when $\theta = 2.7$, and it is illustrative of the results of the simulations.

Table 8 shows the percentiles of the distribution of the net undercount rate for different distributions for the component errors when θ is taken to be 2.7 and the TARO imputation is used. The standard deviation for the posterior distribution was 0.23. In all the cases, a normal distribution is an adequate approximation. The percentiles differed by at most 0.02 for the percentiles between 5 and 95. The 1 and 99 percentiles differed by at most 0.08.

Varying the value of the estimate of θ for the correlation bias did affect the moments of the posterior distribution of the undercount. The variation appears in the mean and in the standard deviation. Table 9 shows the results for the different values of θ , where the distribution for the errors are normal. The case where $\theta = 1$ portrays virtually no correlation bias, while for the other sources of error are present. In the cases where $\theta = 2.1, 2.7$, and 3.7 , all the sources of error are taken into account. The distribution of the undercount shifts to the right as the estimate of θ for the correlation bias increases. The variance also increases as the estimate of θ increases. For all values of θ considered, the bias $B(\hat{U})$ is positive although it decreases as θ increases.

The simulations were conducted with reasonable alternative models for the imputation for unresolved match status. Although there was some variation in the first two moments of the distribution of the net undercount rate, the estimate of net undercount rate in TARO appears robust to missing data. Table 10 illustrates the results of the simulations using models 000 and 111 described in Section 5.7.3. Models 000 and 111 yielded the upper and lower bounds of the undercount estimates under all the reasonable alternative imputation models. The bias in the estimate of the percent net undercount rate ranges from 0.93 to 2.79. In other words, the bias is between 11 percent and 33 percent of the net undercount rate estimate of 8.42. Varying the imputation model has almost no effect on the standard deviation.

Table 10
Posterior Distribution of the Percent
Net Undercount Under Reasonable Alternative
Imputation Models When $\theta = 2.7$

	$E(U)$	St. Dev.	$B(\hat{U})$
TARO	7.24	0.23	1.18
Model 000	7.49	0.23	0.93
Model 111	5.63	0.22	2.79

The total variance of the estimated net undercount rate may be estimated by the sum of the sampling variance and the nonsampling variance. For the case where $\theta = 2.7$, the standard deviation shown in Table 10 for both models 000 and 111 is 0.22 which translate to a non-sampling variance of 0.0005 when all errors are considered. The standard deviation of the estimate of net undercount rate is 0.70 which translates to a sampling variance of 0.49. Therefore, the total variance is 0.0054 and standard error is 0.73. The coefficient of variation of the net undercount rate is 0.083. The nonsampling variance contributes very little to the total variance relative to the contribution by the sampling variance.

7. CONCLUSIONS

When the post-stratification is used in the estimation, the undercount estimate for TARO is 9.02. The post-stratification increased the net undercount rate estimate by 0.6, which is less than one standard deviation of 0.73 from the estimate of 8.42. Although we expect the error in the post-stratified estimate is smaller, the result is consistent with the error analysis.

As we consider all the sources of error in the posterior distribution of the net undercount rate, we do not know the distribution of the correlation-bias parameter θ . Although we could assume a prior distribution for θ , others might disagree. If we were certain that θ is 2.7, then our 95 percent confidence interval for the net undercount rate would be

$$4.77 < U < 9.55.$$

We calculate this by taking the post-stratified estimate 9.02 and adjusting for the two bias estimates in Table 10, 2.79 and 0.93, and two standard deviations, 2×0.73 . We feel this is a conservative estimate since we use two different bias estimates from imputation models 000 and 111. A very conservative 95 percent confidence interval for U for any value of θ between 2.1 and 3.7 is (4.43, 10.32).

We believe the methodology described in this paper is applicable in the 1990 census with appropriate modifications. Areas for further research are nonsampling error estimates for post-strata, a distribution for the correlation-bias parameter, and models for address reporting error.

ACKNOWLEDGEMENTS

The authors wish to thank Aref Dajani for developing the computer software for the simulations and Chris Moriarity for the variance computation. We are also grateful to Kirk Wolter and Howard Hogan for helpful discussions and suggestions during the development of the models. Bruce Spencer thanks the Undercount Research Staff at the Census Bureau for supporting this research under a Joint Statistical Agreement. The authors thank the referees for their helpful comments and suggestions.

APPENDIX

Definition of Balancing Error

The non-linearity of the dual system estimator makes an additive model inadequate for viewing the technical implications of the balancing of the estimated gross overcount and the gross undercount. Therefore, a more appropriated multiplicative model is developed in this section.

Limiting the E-sample and the P-sample search areas affects two parts of the DSE. One effect is a bias in the estimate of the number of erroneous enumerations, \widehat{EE} . The other is a bias in the estimate of the number of people in both the census and the P-sample population, \widehat{M} .

The following definitions are needed for examining the effects of limiting the E-sample and the P-sample search areas in the TARO design on the dual system estimate:

- b = the proportion of the correct census enumerations that are in their P-sample search area.
- g = the ratio of the number of correct census enumerations that are in their E-sample search area to the number that are in their P-sample search area.

The proportion g reflects error in the implementation of the survey committed when the E-sample search area is not equal to the P-sample search area. The way TARO was executed implies $g = 1$. To show what would happen if g does not equal 1, we will carry g through the discussion.

The limiting of the search area causes only a percentage b of the P-sample people who are in both the census and the P-sample population to be designated as matching a census enumeration. Under these circumstances, a systematic bias equal to $(1 - b) N_{11}$ is introduced into the estimation of the number of people in both the census and the P-sample population. Therefore, the observed really estimates bN_{11} .

Likewise, the limiting of the search area causes only a percentage b of the census enumerations to be available to be designated as correct. Then only a percentage g of those, the ones whose search areas are consistent with the proper E-sample search areas, will be designated as correct. Under these circumstances a systematic bias equal to $(1 - bg)N_{1+}$ is introduced into the estimation of the number of distinct people in the census. This bias occurs in the estimation of the number of erroneous enumerations, \widehat{EE} . With this formulation, the observed number of distinct people in the census really estimates bgN_{1+} .

If $g = 1$, no systematic bias is present in the estimation of the dual system estimate because $bgN_{+1}N_{1+} / (bN_{11}) = N_{1+}N_{+1} / N_{11}$.

The error in the estimation of N_{+1} due to the failure to balance may be defined by

- c_b = the error in the number of erroneous enumerations due to the failure to define the E-sample search areas consistently with the P-sample search areas.

The error c_b would be nonzero if g does not equal 1. The ratio g may be greater than or less than 1. The error is given by $c_b = b(g - 1)N_{+1}$.

Measurement

In TARO, c_b was evaluated by testing to confirm that balancing was not an issue and that the design was under control. The percentage of matching enumerations found within the sample block was large, which implies that the design was under control. Since the design was under control, g is assumed to be approximately 1, and c_b is assumed to be negligible.

Estimation

The geocoding appeared to be very good in the TARO test site. However, no formal measurement of the effects of any misassignment on the estimation of \widehat{EE} was conducted. Therefore, g is assumed to be 1, which implies $E(c_b) = 0$ and $\text{Var}(c_b) = 0$.

REFERENCES

- CHILDERS, D., DIFFENDAL, G., HOGAN, H., SCHENKER, N., and WOLTER, K. (1987). The technical feasibility of correcting the 1990 Census. *Proceedings of the Social Statistics Section, American Statistical Association*, 36-45.
- CORBY, C., and MULRY, M. (1988). Memorandum to K.M. Wolter, Subject: Matching Error Pilot Study. Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.
- COWAN, C.D., and MALEC, D.J. (1986). Capture-recapture models when both sources have clustered observations. *Journal of the American Statistical Association*, 81, 347-353.
- DIFFENDAL, G., (1988). The 1986 test of adjustment related operations in Central Los Angeles County. *Survey Methodology*, 14, 71-86.
- ERICKSEN, E.P., and KADANE, J.B. (1985). Estimating the population in a census year: 1980 and beyond. *Journal of the American Statistical Association*, 80, 98-108, 129-131.
- HOGAN, H., and MULRY, M. (1987). Operational standards for determining the accuracy of census results. *Proceedings of the Social Statistics Section, American Statistical Association*, 46-55.
- HOGAN, H., and WOLTER, K. (1988). Measuring Accuracy in a Post Enumeration Survey. *Survey Methodology*, 14, 99-116.
- MORIARITY, C. (1987). STSD 1986 Test Census Memorandum II-12, Subject: Documentation of the Calculation of the Los Angeles Post-Enumeration Survey Block Weights and Dual System Estimate Variances. Statistical Support Division, U.S. Bureau of the Census, Washington, D.C.
- SCHENKER, N. (1988). Handling missing data in the estimation of coverage error for the 1986 census of Central Los Angeles County. *Survey Methodology*, 14, 87-97.
- SEKAR, C.C., and DEMING, W.E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, 44, 101-115.
- SELTZER, W., and ADLAKHA, A. (1974). On the Effect of Errors in the Application of the Chandrasekaran-Deming Techniques. Reprint 14, University of North Carolina, Laboratory for Population Statistics.
- SPENCER, B.D. (1986). Conceptual issues in measuring improvements in population estimates. *Proceedings of the Second Annual Research Conference*, U.S. Bureau of the Census, Washington, D.C., 393-407.

- SPENCER, B.D. (1980). Implications of equity and accuracy for undercount adjustment: a decision - theoretic approach. *Proceedings of the 1980 Conference on Census Undercount*, U.S. Bureau of the Census, Washington, D.C., 204-216.
- WOLTER, K.M. (1986a). Some coverage error models for census data. *Journal of the American Statistical Association*, 81, 338-346.
- WOLTER, K.M. (1986b). A Combined Coverage Error Model for Individuals and Housing Units. SRD Research Report Number Census/ SRD/RR-86/27, Statistical Research Division Report Series, U.S. Bureau of the Census, Washington, D.C.

Representing Local Area Adjustments by Reweighting of Households

ALAN M. ZASLAVSKY¹

ABSTRACT

Suppose that undercount rates in a census have been estimated and that block-level estimates of the undercount have been computed. It may then be desirable to create a new roster of households incorporating the estimated omissions. It is proposed here that such a roster be created by weighting the enumerated households. The household weights are constrained by linear equations representing the desired total counts of persons in each estimation class and the desired total count of households. Weights are then calculated that satisfy the constraints while making the fitted table as close as possible to the raw data. The procedure may be regarded as an extension of the standard “raking” methodology to situations where the constraints do not refer to the margins of a contingency table. Continuous as well as discrete covariates may be used in the adjustment, and it is possible to check directly whether the constraints can be satisfied. Methods are proposed for the use of weighted data for various Census purposes, and for adjustment of covariate information on characteristics of omitted households, such as income, that are not directly considered in undercount estimation.

KEY WORDS: Undercount; Raking; Local-area adjustment; Missing data.

1. HOUSEHOLD-LEVEL ADJUSTMENT BY WEIGHTING

A major research effort has been devoted to methods for estimation of the undercount in the 1990 Census in the United States (National Academy of Sciences 1985). In one of the primary methodologies that has been proposed, a Post Enumeration Survey (PES) would be conducted shortly after the Census in a sample of blocks. The fraction of persons in the PES who were omitted from the Census enumeration yields an estimate of Census undercoverage. Estimates of the undercount would be carried down to some geographical level (possibly the smallest geographical unit used by the Census, the block). These estimates would apply to classes formed on the basis of characteristics of persons, as well as possibly some household or block-level characteristics. The term “class” will be used henceforth to refer to estimation or adjustment classes or cells; the term “block” will refer to the smallest geographical unit for which undercount estimates are calculated. The 1980 Census found approximately one hundred million households in two to four million blocks, depending on the definitions used.

For each block, the outcome of the processes described above would be a vector of estimated undercounts, with S components corresponding to the adjustment, or estimated number of persons omitted from the census in that block, from each of S adjustment classes. The methods by which these estimates are arrived upon are beyond the scope of this paper. However, in our examples we shall assume that for each class within each block there is an undercount rate, expressing estimated omissions as a fraction of enumerated persons in that class and block. In this paper, the term “adjustment” refers to any process which incorporates the estimated undercount into the enumeration. The adjustment classes might be, but would

¹ Alan M. Zaslavsky, Statistics Center, Massachusetts Institute of Technology, Room E40-111, Cambridge, MA 02139, U.S.A. and Harvard University Department of Statistics, Cambridge, MA 02138, U.S.A.

not necessarily be, the same as the post-strata formed in analysis of a Post-Enumeration Program. For forming simple marginal tabulations of persons by characteristics, this information might well be adequate. In particular, small-area counts used for various official and commercial purposes could be calculated from block totals.

However, for some purposes it would be desirable to place the added persons in households. We assume for these purposes that there is also an estimate of the number of omissions of whole households in each block. There might also be information distinguishing omissions of persons within enumerated households from those in omitted households.

If the resulting adjusted records are to be meaningful, the composition of the added households and the relationships of its individual members must be logically consistent and typical of the types of households found in that area. The term “composition” will be used to refer to the number of household members from each adjustment class. Thus, for example, a household consisting of a 20-year old white female head of household, a 75-year-old Chinese male, and a 10-year-old black daughter would not be a very plausible household, even if all of its members were from classes that are well represented in the block. Yet abstractly to describe these patterns and create new households that fit them is a daunting task.

Example 1: *Forming a roster of households.*

Table 1 illustrates part of a census enumeration as it might appear on a microdata tape.

Table 2 represents the same roster, showing how the composition of the households might be summarized if there were only three estimation classes: (1) men over 20 years of age, (2) women over 20 years of age, and (3) children up to 20 years of age.

Table 1
A piece of a sample microdata file

Name	Address	Sex	Age
John Smith	328 Main Street	M	34
Mary Smith	328 Main Street	F	32
Louise Smith	328 Main Street	F	7
Nancy Chen	330 Main Street	F	62
Jorge Ramirez	332 Main Street	M	21
Juan Ramirez	332 Main Street	M	24

Table 2
Microdata file recoded by household, showing
composition of households

Address	Count of persons by class		
	Class 1	Class 2	Class 3
328 Main Street	1	1	1
330 Main Street	0	1	0
332 Main Street	2	0	0

Essentially the same problem arises in many situations in which a household survey must be reweighted to match known marginal totals for various classes of individuals.

The essence of the method proposed in this paper is to assign weights to the households enumerated in the census lists for the block, so that the weighted totals of persons in each adjustment class and the weighted total number of households are precisely equal to the corresponding adjusted totals. Thus, although the weighting changes the proportionate composition of the block, all of the households are real and possess characteristics and relationships that are logically consistent and reasonable for that block. This weighting methodology is similar to the standard raking adjustment, in which the weight applied to counts in a cell of a contingency table is the adjusted count divided by the original count. The household weights are calculated *after* the block totals have been adjusted and will be consistent with those totals. For most Census purposes, the weighted records would be an adequate basis for forming published tables and sampled lists.

This proposal might be contrasted with imputation methods, in which undercounted units are represented by whole units added to the roster. The imputed units may be either persons or households. Although individual persons may be imputed into the block, the problem of fitting these persons into plausible households remains unsolved. Placing them in fictitious "group quarters," as was done in some tests of adjustment procedures, sidesteps this problem at the cost of creating a skewed picture of relationships in the block. Reweighting or imputation of individuals would be appropriate for residents of institutions or group homes, for whom the particular configuration of persons in the dwelling unit has no particular significance.

Another approach to imputation starts with probability models for omissions of households and of persons within households, and draws imputed households from the posterior distribution of the omissions given the enumerated households. This methodology is suited to the multiple imputation approach (Rubin 1987), in which the entire imputation process is repeated several times to represent the variability introduced by the underenumeration. However, in each block roster that is created, totals based on enumerated and imputed households would not necessarily be precisely equal to the desired adjusted totals. In this paper, our concern is with methods that give an *exact* fit to population estimates derived at a preceding stage.

The remaining sections of this paper develop methods for the proposed weighting adjustment. Section 2 gives a mathematical formulation of the objectives of the weighting scheme, while Section 3 explains how to fit the weights. Section 4 explains how to incorporate the distinction between omissions in enumerated and omitted households into the scheme. Section 5 introduces some refinements that improve the robustness of the procedure against the variability of small blocks. Section 6 describes simulation results. Section 7 discusses the use of weighted data for various Census purposes, while Section 8 considers the effects of the weighting adjustment on covariates that are not part of the scheme used in forming the adjustment classes. Finally, Section 9 summarizes some unresolved questions and areas for future research.

2. OBJECTIVES AND MATHEMATICAL FORMULATION OF A WEIGHTING PLAN

It is an essential goal of the proposed plan that the population of the block be assigned to valid household units, so that statistics for which the unit is the household are unambiguously defined. Thus, weights are assigned to *households*; the same weights apply to every *person* within the household.

In order that the counts in the weighted roster be those which are given by the predetermined adjustment, the following constraints must be satisfied:

- (A1) Within each block, the sum of household weights equals the adjusted number of households.
- (A2) Within each adjustment class and each block, the sum of weights for persons equals the adjusted number of persons.

In order that the weighted block roster be as similar as possible to the original block roster, we further require that:

- (B) The weights should be, in some sense, as close to each other as possible.

With unit (or equal) weights, the composition of the block remains unchanged. If the weights are not very unequal, the census composition of the block is nearly preserved by the weighting scheme. To the extent that information about the undercount does not require a drastic revision of our view of the makeup of the block such a drastic revision should be avoided, consistently with good survey practise regarding weights.

We now turn to the mathematical formulation of these criteria. Suppose that in the block under consideration, there are S adjustment classes and I enumerated households, and household i contains C_{is} members from class s . Suppose that H is the desired total number of households in the adjusted roster for the block and D_s is the desired total number of persons in class s . Let $W_i, i = 1, 2, \dots, I$, be the weights corresponding to the households. (A1) requires that

$$\sum_{i=1}^I W_i = H \tag{1}$$

and (A2) requires that

$$\sum_{i=1}^I W_i C_{is} = D_s, s = 1, 2, \dots, S. \tag{2}$$

These constraints can be represented by a matrix equation of the form $AW = B$, where

$$A = \begin{bmatrix} \mathbf{1} \\ C' \end{bmatrix}, B = \begin{bmatrix} H \\ D \end{bmatrix}, W' = [W_1 \ W_2 \ \dots \ W_I] \text{ and } D' = [D_1 \ D_2 \ \dots \ D_S] \tag{3}$$

and $\mathbf{1}$ is a row of 1's.

Objective (B) is represented by selecting some objective function that represents the distance between the weights W and uniform weighting, and minimizing it. We will use the objective function $T = \sum W_i \log (W_i)$. This measure is proportional to the discriminant information (Kullback-Liebler information) of the discrete probability distribution (over households) with relative weights W_i with respect to the probability distribution with equal weights, and is the same objective function that underlies the traditional "raking" (iterative proportional fitting) procedure for adjusting contingency tables (Deming and Stephan 1940; Ireland and Kullback 1968; Oh and Scheuren 1978 have a larger bibliography). Thus, our procedure may be regarded as an extension of raking. Scheuren (1973) applies raking to reweighting of households; Cilke and Wyscarver (1988) reweight to linear constraints but use a different objective function than

those considered here. Methods similar to those presented here were developed independently by Alexander (1987).

In the context of raking, initial counts X are given for cells in a contingency table, and new cell counts Y are calculated to minimize the objective function $\sum Y_i \log (Y_i/X_i)$. Then the weights of the original observations are the ratios $W_i = Y_i/X_i$. In our context, if X_i households happened to have exactly the same composition we could regard them, in the same way, as forming a single entry in the roster with initial count X_i and fit an adjusted count Y_i . However, with a large number of adjustment classes, it would be unusual for several households in the same block to have exactly the same composition. Thus we will not attempt to group households; rather, it is notationally and computationally simpler to list the households separately so that for each enumerated household composition the initial count $X_i = 1$ and $Y_i = W_i$. Aside from this notational difference, the mathematical formulation here differs from that of a raking adjustment only in that the linear constraints do not have the special structure of margins in a contingency table. For brevity in the presentation of examples, we will sometimes include a count on a line to represent that number of identical lines in the roster of households.

In the contingency table setting, raking preserves cross-product ratios of cells, and preserves independence of variables when it holds in the original table. For these reasons, it has been called "structure-preserving estimation" in small-area estimation applications (Purcell 1979; Purcell and Kish 1979). See Section 10.1 for a further discussion of objective functions.

Our procedure differs from raking in that the linear constraints do not necessarily refer to margins in a contingency table. Our methodology includes raking as a special case, as well as the raking generalization of Oh and Scheuren (1978) in which different tables are used to fit each margin. In fact, constraints may be imposed on continuous as well as discrete covariates; applications of this sort are proposed in Section 8.3. Furthermore, the algorithms that are set forth allow direct determination of whether there are in fact any weights that are consistent with all of the given constraints. It is possible then to select constraints that must be relaxed in order to fit weights. These features give these methods potential applicability extending beyond the area of representing undercount.

3. FITTING THE WEIGHTS

The problem before us now is to determine weights satisfying the constraints $AW = B$, $W \geq 0$, minimizing the objective function $T = \sum W_i \log (W_i)$. To make T a continuous function of W , we adopt the usual convention $0 \log 0 = 0$.

We will call any weight vector that satisfies the linear constraints (the equations and the inequalities) a *feasible solution*. As long as there is a constraint on the total weight of the households, the set of feasible solutions is bounded and therefore T assumes a minimum value on it; furthermore, since T is strictly convex, the solution is unique.

The problem of calculating weights then naturally is divided into three tasks: (1) determining whether the linear constraints $AW = B$ are consistent; (2) determining whether there are any feasible solutions; and (3) finding the feasible solution minimizing T . We will suppose that there are I households and p constraints, so A is a $p \times I$ matrix.

Example 2: *Fitting weights.*

Table 3 illustrates the roster of households in a block in which three classes are represented, as in Example 1; we may think of the classes as "men," "women," and

Table 3
A household roster

Line #	Count per household by class			Number of households
	Class 1 (men)	Class 2 (women)	Class 3 (children)	
1	0	1	0	50
2	0	1	1	40
3	1	0	0	40
4	1	0	2	15
5	1	1	0	50
6	1	1	1	60
7	1	1	2	40

Table 4
Adjusted totals

	Raw count	Adjustment rate	Adjusted count
Class 1	205	.05	215
Class 2	240	.03	247
Class 3	210	.04	218
Households	295	.02	301

and “children.” This table may be regarded as a condensed version of a table with 295 lines, each representing one household.

The unadjusted and adjusted counts of households and of persons in each class are found in Table 4. The adjusted counts are calculated by applying the listed adjustment rates and rounding. The method by which the adjusted counts are obtained is immaterial, however, to the rest of the process.

3.1 Consistency of Linear Constraints

As long as the rows of A are independent, the constraints $AW = B$ will be consistent. If any row is dependent on the others, the corresponding constraint is either inconsistent or redundant, depending on the values in B . Dependent rows can be identified by applying the Q - R decomposition to A' . If the corresponding constraints are redundant, they may be deleted without any loss of information; if they are inconsistent, the constraints must be reformulated in some way.

Example 2: (continued).

The A matrix for this example has independent rows, and hence the constraints are consistent.

In Section 5, we consider circumstances in which inconsistent constraints are likely to appear and some methods for dealing with them.

3.2 Existence of Feasible Solutions

Determining the existence of feasible solutions is equivalent to determining an initial feasible solution in a linear programming problem, and the standard algorithms can be used. Suppose

our problem is to find a positive solution W to $AW = B$, where $B \geq 0$. (If the latter condition does not hold it can be made true by reversing the sign of negative elements of B and the corresponding rows in A .) Then create an augmented problem $[A \mid I] [W' \mid Z']' = B$, $W, Z \geq 0$, where I is a $p \times p$ identity matrix and Z is a p element vector variable. This problem automatically has an initial solution $W = 0, Z = B$. Then apply the simplex method (as in Gass (1964) or any other linear programming text) to minimize $\sum Z_i$. If that sum can be reduced to 0, the corresponding W values are a solution to the original problem, while if it cannot, the original problem has no solution.

Example 2: (continued).

A feasible (but not optimal) solution for this example gives total weighted counts of 86, 54, 29, and 132 to the household compositions in lines 2, 3, 5, and 6 respectively of Table 3. It may be verified that these counts yield the desired adjusted totals for households and for individuals in each class.

The problem of infeasibility is similar to that of inconsistency and is also discussed in Section 5.

3.3 Optimizing the Objective Function.

By the method of Lagrange multipliers, the minimizing solution must satisfy the equations $\partial T/\partial W_i = \log W_i + 1 = a_i'\lambda$, where a_i is the i -th column of A and $\lambda' = (\lambda_1, \lambda_2, \dots \lambda_p)$. Then $W_i = \exp(a_i'\lambda - 1)$; thus the model for the weights is log-linear in form, like that for a conventional raking adjustment. λ_s represents the additional log-weight increment associated with a unit increment in the corresponding constraint coefficient a_{is} , *i.e.* adding an additional household member from adjustment class s to the household.

We can solve for λ by Newton's method to satisfy $AW = B$. The iterative scheme we use is

$$\lambda^{(t+1)} = \lambda^{(t)} - (AW^*A')^{-1} (AW - B), \tag{4}$$

where W^* is the matrix with the elements of $W = W(\lambda^{(t)})$ on the diagonal. A good starting value for λ is $\lambda^{(0)} = (AA')^{-1}B$, which can be derived from a linear approximation around equal starting weights. A cyclic descent procedure for solving these equations, which is a generalization of iterative proportional fitting, is described in Section 10.2.

Example 2: (continued).

The weights per household and total weighted counts (weight times raw count) for each line in Table 3 are shown in Table 5. No household is upweighted by more than 8% or downweighted by more than 5%.

Table 5
Optimal weights for Example 2

Line #	Weight	Weighted counts
1	0.9554	47.77
2	0.9557	38.23
3	0.9816	39.27
4	0.9823	14.73
5	1.0730	53.65
6	1.0734	64.40
7	1.0737	42.95

4. WHOLE-AND WITHIN-HOUSEHOLD ADJUSTMENTS

We now consider the distinction between within-household adjustments (that is, adjustments for omissions of persons within enumerated households) and whole-household adjustments (that is, adjustments for omissions of whole households). This distinction has previously been made for purposes of analysing the causes of undercount (Fay 1986). Our concern here is to use it to more accurately represent the undercount by an adjustment.

Within-household adjustments do not involve adding any households to the roster, but only shifting weight between households to increase the weighted totals of persons in the various classes. That is, households with few or no persons in a particular class are downweighted and those with many are upweighted, so that the total household weight remains constant. Thus, in this portion of the adjustment, some households will inevitably have their weights reduced. Whole-household adjustments, on the other hand, correspond to households that were omitted entirely from the census. These adjustments do not reflect on the accuracy of the enumerated households; thus they should be represented by adding households to the roster without taking weight away from the households that were enumerated.

We propose to separate these two portions of the adjustment. One set of constraints represents the within-household adjustment. The total household weights are here constrained to equal the enumerated count of households, while the total weights assigned to persons in each class are constrained to equal the enumerated count in that class plus the within-household adjustment for that class. $AW_1 = B_1$ where B_1 consists of the *enumerated* household count and the counts of persons by class adjusted for *within-household* undercount.

A second set of constraints represents the whole-household adjustment. The total household weights are here constrained to equal the estimated omitted households, and the total person weights in each class are constrained to equal the estimated omitted persons in those households. $AW_2 = B_2$ where B_2 consists of the count of added households and the counts of added persons by class for the adjustment for *whole-household* undercount.

After fitting two sets of weights corresponding to the two sets of constraints, the two weights for each household are added to obtain weights that incorporate both parts of the adjustment ($W = W_1 + W_2$). The distinction between whole- and within-household adjustments contains information which may lead to a different set of adjusted weights than would be calculated if the adjustments were combined, as is illustrated in Example 3. However, if this distinction is not made in the estimation of the undercount, an adjustment can still be calculated in a single step.

Example 3: *adjustments for whole-household omissions.*

Suppose there are only two adjustment classes, and a hypothetical block has the composition described in the first three columns of Table 6.

Suppose now that to the 30,010 households enumerated, we must add 231 persons each in Class 1 and Class 2, and 121 households. The last three columns of Table 6 show the adjusted counts under alternative assumptions: (1) the omitted persons may belong to any household, enumerated or omitted, and (2) all of the omitted persons were in the omitted households.

When the omitted persons could have been in any household, the algorithm downweights the households with only one person from each class (1,1) and upweights households with two from one class and one from the other (1,2 and 2,1). While the households with two persons from each class are substantially upweighted (by a factor of 1.354), only a small portion of the added persons appear in those households since

Table 6
Hypothetical raw and adjusted household counts for Example 3

Household composition		Raw count (number of households)	(1) Omitted persons in any household	(2) Omitted persons in omitted households only	
Class 1 persons	Class 2 persons		Adjusted counts	Counts of omitted households	Adjusted totals, omitted and enumerated households
1	1	10,000	9904.54	.01	10,000.01
1	2	10,000	10106.46	10.99	10,010.99
2	1	10,000	10106.46	10.99	10,010.99
2	2	10	13.54	99.01	109.01

the original count for that composition is so small.

When the omitted persons appear only in the omitted households, weights are calculated first to fit $231 \times 2 = 462$ persons into 121 additional households, and then these weights are added to the unit weights in the raw counts. While no household composition is downweighted, the (2,2) households are upweighted extremely (by a factor of 10.901). In fact, it is mathematically impossible to accommodate 462 persons in 121 households of two to four persons each without having at least 99 households with 4 members. Thus, the information that the added persons (or some known fraction of them) belong in the omitted households substantially changes our view of the appropriate adjustment.

5. FEASIBILITY OF CONSTRAINTS

In the preceding sections we have assumed that feasible solutions exist to the constrained optimization problem. Here we will consider situations in which the solutions will not exist or will be unsatisfactory, and some alternative methods to deal with these situations.

5.1 When Will Constraints be Non-feasible?

There are three ways in which the constraints may fail to allow of satisfactory solutions: (1) when the constraints are actually inconsistent, (2) when the constraints are consistent but there are no positive weights that satisfy them, and (3) when there is a feasible solution but it involves an extreme adjustment to some household weights. The issues associated with these three failure modes are fairly similar.

One could write down constraints that are intrinsically inconsistent, for example that all classes of men are adjusted upward by 2% while men in total are adjusted upward by 4%. In our procedure each constraint applies to the number of persons in a distinct adjustment class and so there are no inconsistencies of this sort. However, a contingent inconsistency is still possible, that is to say one that depends on the particular collection of household compositions that appears in a block. The following are examples of contingent inconsistency, infeasibility, or unsatisfactory weights:

- (1) Proposed undercount estimation methods envision defining over 100 adjustment classes. In a small but diverse block the number of classes represented might be larger than the

number of households; hence the number of constraints would be larger than the number of weights to be fitted. An inconsistency is then almost inevitable.

- (2) If all households in the block have exactly the same number of members from a particular adjustment class (*e.g.* every household has one young Hispanic girl), then the number of members of this class represented is unaffected by the distribution of weights.
- (3) The adjustment of the number of households may be too large or small to accommodate the adjustment of persons in some class. This may represent a failure of the model for adjustment of the number of households. For example, suppose that the number of men to be added by the whole-household adjustment is greater than the number of households to be added, but no household in the block has more than one man. The constraints then might be consistent but infeasible, since they could be satisfied only by assigning negative weights to some households without men.
- (4) The block may have had omission rates atypical of blocks in the PES on which omission rates were estimated. For example, suppose that in most blocks (including most of the PES sample blocks), adult males with certain characteristics tend to be heavily undercounted, but the block being adjusted is atypical in having adult males of this class present in most households and well counted. The class undercount estimate might lead to an extreme upward adjustment that could not be accommodated within the existing households.
- (5) Some adjustment may require giving substantial additional weight to households containing persons from a combination of adjustment classes that appears in only one household, so that household receives an extreme weight. In this case the problem is feasible but the solution is not very satisfactory.

Problems of infeasibility may also arise where the difficulty cannot be so easily traced to a particular inconsistency in the adjustment.

5.2 Making the Constraints Feasible

Regardless of the stage of the fitting procedure at which the infeasibility is discovered, several methods are available to relax the constraints and make them feasible. In this section, we survey several such methods, drawing out both the intuitive logic of each choice and the computational methods required.

5.2.1 Methods Based on Dropping Rows (constraints) of A

When checking for consistency of constraints, some rows may be found to be linearly dependent on the previous rows and hence either redundant or inconsistent. If these rows are simply dropped from the A matrix, a consistent set of constraints is obtained; thus, no further computational effort is required.

If the constraints are arranged in sequence from the most important to the least important, then the less important constraints will be dropped when they are inconsistent with the more important ones. This ordering makes the most sense if the original constraints on distinct adjustment classes (defined by a multi-way classification of the population) are reframed in an ANOVA-like manner as constraints on total population ("grand mean"), classes defined by one classification variable ("main effects"), and classes defined by interactions. For example, if there are ten adjustment classes defined by two sexes and five age ranges, the reframed constraints in order of importance might be: total population (1 constraint), population by sex (1 more constraint), population by age (4 more constraints), age-sex interactions (the remaining 4 constraints). The 4 age constraints could be further broken down as old-vs.-young (1 constraint) and 3 further constraints within those larger groups.

A similar procedure can be applied at the stage of checking feasibility of the constraints. If it is not possible to make all of the $Z_i = 0$, the objective function in the linear programming problem can be modified to be $\sum c_i Z_i$, with the coefficients $c_i > 0$ corresponding to the most important constraints made larger. Then a maximal set of feasible constraints can be identified, and the remaining constraints dropped.

The outcome of this procedure would be weights that give the correct block totals on the coarser classifications of persons, while failing to be correct on all cross-tabulations.

5.2.2 Methods Based on Adding Columns (households) to A

When constraints are only contingently infeasible (in the previous sense that infeasibility depends on the particular set of household compositions in the block), they become feasible when households are added that have the required composition. The simplest application of this principle is to work at a higher level of geographical aggregation than a block. A few adjacent blocks may be combined when problems arise in fitting, or the entire roster may be grouped at, for example, the enumeration district level before weighting. The larger the unit, the broader the range of household compositions that will be represented and the less likely that problems of infeasibility will arise.

A more sophisticated procedure would use a hot-deck of households from adjacent “donor” blocks to enrich the pool of households to which weight can be assigned. Computational simplicity is important here since it may be necessary to scan through a long list of households to find the one or ones which will make the constraints feasible. In the consistency-checking stage, if row j of A is dependent on the previous rows, then if the column for the added household is independent of the columns of A (with regard only to the first j rows), row j of the augmented A will be independent. In the stage of checking for feasibility, if the algorithm halts because no reduction can be made in the objective function $\sum Z_i$, the search for basic columns can be extended to columns corresponding to households in the hot deck. Finally, if some household’s fitted weight is extremely high, the hot deck can be scanned for other households that would also receive high weights with the current values of λ (that is, columns a such that $a' \lambda$ is large). If these are added to the block they will draw off some of the weight from the overweighted households when the weights are refitted, since they are likely to also have members in the same adjustment classes.

The intuition behind this method is that the household compositions that are enumerated in a block are only a sample of those which actually could have appeared there had the enumeration been complete. The observed distribution of household compositions is smoothed by mixing it with the distribution for adjacent blocks, which contain households that are also typical for that area. Thus, conceptually this method is related to Bayesian smoothing methods that improve estimation of some quantity for one unit by borrowing strength from its distribution in similar units. This Bayesian rationale is developed in terms of a block-level random-effects model by Zaslavsky (1989).

The donor blocks could be chosen by a sequential hot deck procedure; then, the donor blocks would tend to be geographically close to the adjustment block and no particular set of blocks would have undue influence on the entire census. By detailed stratification of blocks, the donor blocks could be selected to be similar to the block being adjusted on characteristics such as mean income, types of housing units, and racial balance.

5.2.3 Combined Methods

The two types of methods outlined above can be combined by an appropriate reframing of constraints. The principle here is to satisfy *all* constraints in the larger geographical units,

while satisfying only the more important constraints in the smaller units. This type of compromise may make it possible to get a fairly good fit to the desired distribution without having to add additional records to the roster.

Suppose that the A matrices for several blocks have been reframed similarly as sequences of rows representing main and interaction constraints. Then a single large A matrix representing all of the constraints can be formed. The rows for the more important constraints can be kept separate, while rows for subsidiary constraints can be combined across blocks. For example, suppose there are ten adjustment classes, defined by sex (2 levels) and age (5 levels), and two blocks. Altogether there are eleven constraints (one for number of households and one for each adjustment class) in each block. If these are combined into a single matrix, keeping main effects and two-way interactions, the constraints are: block household counts (2 constraints), block populations (2 constraints), sex (1 constraint), age (4 constraints), block \times sex interaction (1 constraint), block \times age interaction (4 constraints), and sex \times age interaction (4 constraints) in the combined blocks. Here 4 constraints have been eliminated (block \times sex \times age interaction); in a more realistic problem with more blocks, classification variables, and levels, the reduction would be much greater.

6. SIMULATION RESULTS

Simulations were performed to answer two classes of questions:

- (1) The first set of questions is concerned with evaluation of the success of the algorithm in terms of its own constraints and objectives. Does the reweighting algorithm give an answer? In real problems, is there a solution to the weighting constraints? How much do the weights vary? Is the amount of computation required within reasonable limits?

To answer these questions, "feasibility simulations" were performed in which the weighting algorithm was applied to simulated blocks made up of real households, using real adjustment rates. This procedure thus closely parallels the practical application of the algorithm.

- (2) The second set of questions is concerned with evaluation of the success of the algorithm in improving the quality of inferences based on a micro-data set: does the weighted micro-data set more accurately describe the real world than the raw, unweighted data?

To answer these questions, simulated blocks made up of real households were drawn, representing the true (but unobserved) compositions of households in blocks. For each "true" block, omissions were imposed using real estimated undercount rates and a plausible model for the distribution of undercount among households. The weighting algorithm was applied to the "enumerated" blocks generated in this way. Summary statistics describing household composition were calculated for the simulated "true" blocks and for the simulated observed blocks with undercount, both unweighted and weighted for undercount adjustment. The goal of these "inference simulations" was to determine whether the reweighting brought the statistics closer to their values in the "true" blocks; in other words, did reweighting correct the biases caused by the undercount?

The source of households for all simulations was the 1% "B" Public Use Microdata Sample (PUMS) from the 1980 Census (Bureau of the Census 1985). Households were extracted from sections of Los Angeles County, California that include the site of the Test of Adjustment Related Operations (TARO) of the 1986 Test Census.

Undercount rates were those calculated from the 1986 TARO (Diffendal 1988, Table 7) for adjustment classes defined by sex, age (five levels), race (Hispanic, Asian, or "other race"),

and tenure (owner or renter). Adjustment factors calculated from the given undercount rates ranged from 0.982 to 1.211.

Each household was coded as a vector of counts representing the number of individuals in that household from each of the 60 adjustment classes.

Further details on the simulation procedures and on a larger set of simulations are in Zaslavsky (1989).

6.1 Feasibility Simulations

For each of four block sizes (20, 50, 100, and 200 households), 50 simulated blocks were drawn from the full sample and 50 were drawn from only those households with no Asian members. For each block, simulations were attempted using two levels of the household adjustment rate (the factor by which the number of households in the block is adjusted).

The algorithms of Section 3 were applied. To recapitulate, the linear constraints were checked first for consistency, and then for feasibility (existence of a positive solution); finally, weights were calculated using Newton's method. As no data were available distinguishing within-household and whole-household omissions, no effort was made to separate them in these or other simulations.

The results of these simulations are summarized in Table 7.

Consistency and feasibility:

The columns headed "incons", "infeas", and "OK" represent the number of simulated blocks (out of the 50 trials) in each simulation that fell into each of the following categories respectively: (1) the constraints were inconsistent (could not be satisfied by any weights), (2) the constraints were consistent but not feasible (could not be satisfied by any positive weights), or (3) the constraints were both consistent and feasible.

In the "non-Asian" simulations there are 41 constraints to be satisfied (some of which may be trivial, *i.e.* when the corresponding adjustment classes are unrepresented in the block). Thus with 20-household blocks, the constraints were never consistent; with 50-household blocks, the constraints were sometimes consistent and then usually feasible. The constraints were usually feasible in 100-household blocks, and always in 200-household blocks.

The numbered columns at the right represent the order of the simplest marginal constraint that could not be satisfied, in the sense of the hierarchical reparametrization in Section 5.2.1. Thus, column (1) indicates the number of simulated blocks for which a "main effect" constraint (marginal total of persons classified by one stratifying variable) could not be satisfied, column (2) indicates the number of trials for which a two-way interaction constraint could not be satisfied, etc. Even when the constraints were inconsistent with 50- or 100- household blocks, the main-effect constraints and often the two-way or even three-way interactions were feasible. This suggests that pooling of blocks for higher-order interactions, as described in Section 5.2.3, might be a successful strategy for dealing with problems of infeasibility.

The results were less encouraging for simulations using the full samples. Even with 200-household blocks, only rarely were the constraints consistent and feasible. With increasing block size the lower-order constraints were more likely to be feasible. This is explained by the small number of households with Asian members (approximately 5% in each sample). Out of 200 households, the expected number of Asian households would be about 10, an insufficient number to satisfy the 20 possible constraints for the Asian adjustment classes. Such a situation in which some groups of adjustment classes are poorly represented in a certain region or in particular blocks would surely not be unusual in practice. This would require pooling of blocks on a large scale for the corresponding constraints, while the constraints for the better-

Table 7
Feasibility simulation results

Non-Asian Households													
size	HH	rate	incons	infeas	OK	maxW	minW	varW	iters	(1)	(2)	(3)	(4)
10		1.00	50	0	0	NA	NA	NA	NA	22	28	0	0
10		1.05	50	0	0	NA	NA	NA	NA	8	42	0	0
20		1.00	50	0	0	NA	NA	NA	NA	0	50	0	0
20		1.05	50	0	0	NA	NA	NA	NA	0	50	0	0
50		1.00	47	1	2	1.921	0.200	0.142	3.00	0	3	37	8
50		1.05	47	0	3	1.550	0.620	0.036	1.33	0	3	36	8
100		1.00	10	0	40	2.068	0.429	0.088	2.03	0	0	8	2
100		1.05	10	0	40	1.573	0.753	0.020	1.90	0	0	8	2
200		1.00	0	0	50	2.434	0.543	0.063	2.18	0	0	0	0
200		1.05	0	0	50	1.749	0.821	0.015	2.00	0	0	0	0
Full Sample													
size	HH	rate	incons	infeas	OK	maxW	minW	varW	iters	(1)	(2)	(3)	(4)
100		1.00	49	0	1	--	--	--	--	0	34	15	0
200		1.00	49	0	1	--	--	--	--	0	2	43	4

represented classes might be satisfied on a smaller scale.

Weights:

The maximum and minimum household weights and the variance of the weights were calculated for each simulated block for which the constraints were consistent and feasible. For each simulation condition, the average value of these quantities (across simulated blocks) is displayed under the heads “maxW”, “minW”, and “varW.” The following observations characterize some of the effects of the simulation design factors on the fitted weights.

- (1) For simulations with household count adjustment factor of 1.05, in every case, the average variance of the weights was smaller, and the average of the minimum weights and of the maximum weights were closer to unity, than with household adjustment factor 1. This is intuitively reasonable since almost all class adjustment factors exceed 1, and it requires a more extreme adjustment to add individuals to existing households than to add individuals and households to accommodate them. For example, if the adjustment factors for households and for every adjustment class are all equal, every household would be upweighted equally.
- (2) Fixing other factors, the variance of the weights becomes smaller as the number of households per block increases. Again, this is intuitively reasonable because the pool of households is richer in a larger block; the probability of finding exactly the households needed to represent undercounted individuals is higher. The trends for the extreme weights are less clear-cut than for variances; here, the narrowing of the variance is offset by the larger sample over which the extreme is calculated in the larger blocks.

- (3) The average variances for simulations with 200-household blocks were at most .063. Thus the reweighting is generally not extreme.

Computational costs:

The mean number of Newton steps required to fit the weights (from the starting values given in Section 3.3), shown under the heading "iters", is usually about two. These iterations were sufficient to satisfy all constraints with error no greater than .001. Using this information, a rough estimate can be given of the number of floating point operations required to apply the algorithm. Computational costs of the modified raking algorithm are discussed in Section 10.2.

Assume that blocks are of sufficient size that it is not necessary to check consistency and feasibility of the constraints in every case (but perhaps only when the weight fitting does not succeed in a few steps). Then the key calculation is fitting the weights. For production runs, data structures and programs should be devised which take advantage of the sparseness of the A matrix (due to the fact that only a few classes are represented in each household). Then if S_1 is the total number of nonzero entries in A and S_2 is the sum (through the block) of the squares of the number of nonzero entries for each household, each Newton step requires about $5S_1/2 + S_2/2$ multiplications (plus a term independent of the number of households per block). In the samples studied here, $S_2 \approx 5S_1$; S_1 is bounded by the total population of the block. Thus the bound on the number of multiplications is approximately $15 \times$ population total (counting the start as an iteration); the number of additions is comparable.

In an era in which even microcomputers have megaflop arithmetic capability, 8×10^9 floating point operations to reweight an entire census does not seem unreasonable. The calculation of weights might well take less computer resources than the "bookkeeping" data processing required in any method of incorporating undercount. Of course, if the procedure were applied to a sampled database, as in forming a public-use sample, the costs would be reduced correspondingly.

6.2 Inference Simulations

For the inference simulations, pseudo-blocks of 50 households each with only Hispanic members were drawn. These were treated as if they represented true blocks. Then simulated omissions were imposed on these households, assuming that each member was (independently) omitted with probability equal to the undercount rate from Diffendal (1988), with two negative undercount rates truncated to 0.

The entire distribution of the "enumerated" block was represented by including in the pseudo-Census roster the true composition and the possible compositions obtained by omission of one or more household members, each weighted by its probability under the model.

The pseudo-Census roster with undercount was then reweighted to the original pseudo-block totals for number of households and of individuals in each adjustment class. Both the pseudo-Census roster and the reweighted roster were compared to the original pseudo-block.

The purpose of organizing the simulation in this manner was to remove variability due to randomness in the rate of omissions in a block (around the mean undercount rate) and in the distribution of the omissions among the households in the block. Furthermore, feasibility is guaranteed because the original households are always included (with weights) in the pseudo-Census roster. One way of regarding this setup is that each simulated block represents a very large population in which observed undercount rates and the distribution of observed compositions approach their expectations.

Several sets of statistics were used in evaluation of the reweighting procedure. These were all chosen because they summarized household characteristics that are not functions of the populations by adjustment class. The first set was the distribution of sizes (number of members) of households. Note that the mean number of persons per household, like any function of the class totals and household count, will automatically be adjusted to the correct (pre-undercount) values; the distribution of sizes, however, is not controlled by the adjustment procedure.

The second set of statistics was the distribution of number of *adult* (over 14 years old) members in households with one or more *children* (up to 14 years old). In this case, the mean is not automatically adjusted to the correct value, since it depends on the joint distribution of counts from different classes within households as well as on marginal totals.

The last two sets of statistics were the distribution of the age group (coded from 1 to 5 as in the formation of the adjustment classes) of the *oldest male* in the household (coded 0 if no male is present), and the same distribution for households with one or more children. Again, neither the distribution nor its mean are directly constrained to their true values.

The results of these simulations are summarized in Table 8. Because almost all of the differences noted here are highly significant (relative to between-pseudo-block variances of the differences), standard errors are not shown in the tables. The lines of each table are labelled "true" (for the original pseudo-blocks), "enum" (for the simulated enumerated blocks, *i.e.* after omissions due to undercount), and "adjust" (enumerated blocks after adjustment for undercount). Every column except the means should be read as a percentage of households in the block.

Table 8
Inference simulation results

Size distribution							
	size 1	size 2	size 3	size 4	size 5 +	mean	
true	7.240	16.200	20.240	22.600	33.720	3.971	
enum	10.349	19.631	21.772	20.690	27.558	3.632	
adjust	7.372	16.421	20.596	21.392	34.219	3.971	
Size distribution (number of adults) for households with children							
	size 0	size 1	size 2	size 3	size 4	size 5 +	mean
true	0.000	6.925	58.404	17.214	9.125	8.332	2.585
enum	1.736	18.309	49.874	15.965	7.677	6.439	2.323
adjust	0.924	13.277	48.557	18.223	9.810	9.209	2.562
Age of oldest male (by five age classifications)							
	none	age 1	age 2	age 3	age 4	age 5	mean
true	7.080	4.000	28.680	33.800	21.960	4.480	2.730
enum	9.981	7.388	26.296	30.972	21.160	4.203	2.585
adjust	7.853	5.989	26.307	33.439	21.931	4.480	2.690
Age of oldest male (by five age classifications) for households with children							
	none	age 1	age 2	age 3	age 4	age 5	mean
true	3.602	6.214	30.744	42.649	15.843	0.949	2.638
enum	5.809	11.723	27.321	39.096	15.158	0.894	2.488
adjust	4.272	9.069	27.242	42.038	16.418	0.962	2.601

Household size distribution was biased downwards in the enumerated blocks. As well as correcting the mean, adjustment brought the estimated percentage for every size substantially closer to the true percentage.

The distribution of number of adults in households with children was also biased downwards. The majority of these households had contained two adults, so this size category was most understated by the enumerated statistics. Due to the log-linear structure of the adjustment, however, the most extreme adjustments were made to the largest and smallest households. Thus, the highest size categories were slightly overadjusted and intermediate categories were underadjusted; the "size 2" category was adjusted a small amount in the wrong direction. Nonetheless, the mean of the adjusted distribution was much closer to the "true" value than the adjusted mean was.

The story is similar for the distributions of age of oldest male. Although these statistics are only indirectly related to the counts by class, in almost every case the adjusted distributions and means are closer to the "truth" than are the unadjusted distributions and means.

In summary, these simulations suggest that these weighting adjustments can improve estimates of measures of household structure as well as the aggregate counts for which they were intended. However, reweighting does not provide accurate adjustments with certain configurations of the data, such as the many households with two adults noted above; to deal with these situations may require a model-based imputation method such as that outlined by Zaslavsky (1989).

7. THE USE OF WEIGHTED DATA

The product of the methods of the preceding sections would be a census roster in which households have weights, persons in households have weights adopted from their households, and institutionalized persons have individually assigned weights. This section outlines the use of these rosters for various Census purposes.

7.1 Formation of Tables of Counts

As with any data set of weighted observations, the sum of weights replaces the simple count of observations in forming tables. The only problem created by the use of weights is that of obtaining integer entries in the tables. This problem arises even before the calculation of household weights: when the estimated omissions are calculated, the counts in each class will not in general be integers.

If the adjusted totals by class are rounded to be integers, any table that aggregates classes (for example, a count of adult males that is a sum of counts of adult males from different classes) will also contain integers, since it must be consistent with those totals. For tables that are not based on those totals, summing the weights in a particular group may not necessarily generate integer counts. For example, if a class combines women of ages 20-40, a sum of weights for women aged 20-30 would not necessarily be an integer. In any case, it seems unlikely that all class weights would be rounded since this might well lose the entire adjustment to roundoff error. However, it should be possible to use existing Census Bureau integerizing methods ("controlled rounding") to deal with these problems, especially where non-disclosure requires that published counts be rounded anyway (Cox *et al.* 1986; Cox 1987).

7.2 Formation of tables of sums and means

Generally, sums (of continuous quantities) and means are not expected to be integers, so

the issue of rounding does not arise. Also, tables based on long-form information are already derived from a sample so an additional source of weights should not change the process much. A deeper issue is that of the values of non-classification covariates to be assigned to households that are "weighted in" to the census; this is discussed in Section 8.

7.3 Public Use Samples

The public use tapes are a sample of census records that are released for further analysis by consumers of census data.

To generate these samples from weighted census rosters requires only that the sampling procedure be modified slightly to make sampling probabilities proportional to weights. Even on the 5% tape (the highest sampling rate), the weighted sampling probabilities should be smaller than 1. Once these tapes are produced, the user would not have to be aware of the adjustment and weighting process that had gone into generating them.

The public use tapes are the source of data for many of the more complicated analyses by sociologists, economists, planners, etc. in which the details of household composition, as well as counts of persons, are of importance. It is important that these tapes could be generated easily and used like raw census data.

As a service to those users of the public use tapes who wish to check the sensitivity of their analyses to the undercount adjustment, the tape should include factors (the inverse of the adjustment weights attached to the household records in the original census rosters) that would allow the user to reconstruct the equivalent of the unadjusted census.

8. ADJUSTMENT OF COVARIATES THAT ARE NOT USED IN CLASSIFICATION

The methods described above guarantee that weighted block totals by variables used in classification, such as sex, race, and age group, will equal the adjusted block totals. However, these lists will also be used to accumulate totals or counts for variables such as income and education that might not be used in the classification scheme. This section will consider the effect of these adjustment methods on such statistics. For concreteness of exposition, income will be used as the main example. Income is an important non-classification variable; some research suggests that revenue allocation programs may be most affected by errors in measurement of income. (National Academy of Sciences 1985).

In general, there are two possible sources of bias in the estimation of a non-classification covariate: (1) bias in adjustment of household composition, and (2) systematic differences between fully enumerated households and households with similar composition that are omitted (entirely or in part). However, if we have an estimate of mean income for the block, we can make the weighted mean for households in the block equal the estimated (adjusted) mean in much the same manner we make the weighted counts of individuals in the block equal the estimated (adjusted) counts.

8.1 Household Composition Bias

In this section we will assume that the average income level associated with a certain household composition is the same for fully enumerated households and those which are partly or wholly omitted from the enumeration. In other words, we consider here the case in which omission is noninformative for income.

Suppose that household income is a sum of independent contributions from persons of each class in the household (*i.e.* suppose that the contribution to income from persons in each class are independent of what other members are in the household). Then weighted household income totals would be an unbiased estimate of the true income totals (when adjustment rates are correct), since the sum of incomes would be a linear function of class counts for the block. However, under the more realistic assumption that linearity does not hold, misallocation of persons between households (and corresponding misrepresentation of household composition in the adjustment) could lead to bias in income estimates. Thus, for example, the average income of households with two children might not be the mean of the average income of one-child and three-child households (with the same composition of adult members). Then the weighting procedure might introduce the correct number of children but if, on the average, too many (compared to the truth) two-child households were created relative to one- and three-child households, estimates of household income would be biased.

Our procedure tends to fit weights that make the "adjusted-in" households similar in composition to those that are common in the enumeration. However, the adjustment is described only by adjustment class totals, which do not carry detailed information on the composition of the omitted households. Thus, if certain household compositions are disproportionately undercounted they may be underrepresented in the weighted lists, and if these compositions are associated, for example, with lower incomes, the total income estimates will be biased upwards.

This is essentially a problem of potential lack of fit of the model used in adjustment to the patterns in the data. The most severe biases might appear in statistics that refer specifically to household composition, such as the number of single-parent families.

If composition bias were found to be a serious problem, one approach to controlling it would be to augment the class adjustment rates with additional information that describes the joint omissions of persons from different classes (or grouped classes).

8.2 Response Bias

It is not unreasonable to think that, of a group of households with the same composition, those which are missed in the census will differ systematically in some characteristics from those that are enumerated. In other words, omission may be a form of nonignorable nonresponse. For example, households with lower incomes and educational levels may be more likely to be missed altogether, or to omit some members from their roster; income and education are not classification variables and therefore are not directly adjusted.

Whole-household adjustments are represented in the proposed methods by upweighting households, preserving the values of all covariates. The implicit assumption is that the omitted households do not differ on these covariates from enumerated households with similar composition. There is no information available in the block being adjusted to contradict this assumption. However, it should be possible to collect information in the PES on the differences between enumerated and missed households, which could be incorporated into the adjustment. For example, the income of wholly omitted households might be related to the mean income of enumerated households with the same composition by a linear regression; then the added (weighted-in) households could be imputed the income obtained by applying the linear regression function to the income of the enumerated donor household. Little and Rubin (1987) discuss relevant methods for missing data problems with informative nonresponse. Another approach that is integrated with the weighting adjustment methodology is described in the next section.

Within-household adjustments are represented by downweighting a household with certain enumerated characteristics and upweighting another household that contains an additional

member or members. In the absence of further adjustment, the characteristics of the upweighted household, rather than those of the enumerated household from which the weight was taken, will apply to the “weighted-in” component.

This poses problems that cannot be resolved without collecting some data (from a subsample of the PES). For example, if a *child* were omitted from the household roster, there is no reason to think this would lead to misreporting of income. If households with more children had a higher mean income than those with fewer children, then the weighting would tend to over-estimate mean incomes.

If an *adult* were omitted from the roster, this might also mean that the same adult’s income (if any) would be left out of the reported household income. It is plausible that the mean unreported income in this situation would be positive but less than the mean income of the corresponding adults in households where all adult members appear on the roster. For a stereotypical example, consider a family on public assistance that does not report an adult male member, whose income would otherwise be deducted from the assistance level, and whose residence is somewhat inconsistent. That member’s income is likely to be less than that of a permanently resident adult male in a family that does not depend on public assistance. Thus, neither the income of the enumerated household nor that of the “weighted-up” household would be an accurate imputation for the adjusted household.

No direct correspondence is established between households that are down-weighted and those that receive additional weight. Thus an unadjusted income cannot be carried over directly from the enumerated household to the “weighted up” household. However, with some research comparing the incomes of enumerated and missed households, the incomes of down-weighted households could be used in adjusting incomes. For example, the mean household income of the reweighted block could be constrained to be equal to that of the block before adjustment.

8.3 Weighting Adjustment of Non-classification Characteristics

Suppose that adjusted summary information (by block) is available on some characteristics of households other than counts of individuals by adjustment class. For example, we might have an adjusted estimate of mean income or of the proportion of single-parent families, possibly from a regression model. As long as the summary statistic can be represented as a weighted sum of covariate values for each household, then conformity to the desired adjusted value can be imposed by a linear constraint on weights which can be made part of the weighting adjustment methodology of this paper. Thus, in the income example, we would constrain the weighted sum of incomes to equal the product of the number of households and the adjusted mean income. To adjust the proportion of single-parent families, we would constrain the weighted sum of 0-1 indicators for that status to the desired total count.

8.4 Summary and Implications

The methodology proposed will upweight households, and without further consideration of possible biases, will carry along the characteristics of the upweighted households. If the size of the adjustment and the biases introduced in household characteristics are both of small order, the overall bias in estimated block characteristics will be of second order and should not be a major problem. Some simple regression adjustments might make it possible to reduce the biases by an additional order of magnitude.

9. SUGGESTIONS FOR FUTURE RESEARCH AND DEVELOPMENT OF METHODOLOGY

This section summarizes a number of suggestions for implementation and further development of this adjustment methodology.

9.1 Post Enumeration Survey (PES) Data-gathering and Statistical Modeling

Omissions of persons in enumerated and omitted households should be distinguished in the PES and the two omission rates modeled separately for each adjustment class. Rates of omissions of whole households should also be modeled (Section 4). A variety of measures (as in Section 6.2) could be used to compare the composition of "weighted-in" households to that of omitted households found in the PES; if research found that "composition bias" was a significant problem, higher-order statistics should be developed (Section 8.1). A sample of PES households that were omitted in the Census should be administered the long form, so that the relationship between omission and covariates such as income and education could be modeled for the adjustment (Sections 8.2, 8.3).

9.2 Feasibility of Adjustments

The methods of Section 5 should be tested and compared using PES data.

9.3 Multiple Imputation

Although the procedures proposed in this paper operate deterministically, there are a number of sources of uncertainty in statistics based on the weighted records. These include: uncertainty in estimation of undercount rates; variability in class undercount rates from block to block around the national mean; binomial variability in the actual number of omitted households or individuals around the expected number given the undercount rate; uncertainty regarding differences between covariate values for omitted households and for enumerated households that are weighted up to replace them.

For research uses, files could be prepared that would represent all of these forms of uncertainty by multiple imputation (Rubin 1987). Two or more versions of the reweighted data set could be represented by including several sets of weights on the file. Researchers could repeat their analyses using each set of weights in turn. The variability among the statistics calculated on the different versions gives an estimate of the variability introduced by the process of undercount adjustment. Zaslavsky (1989) discusses procedures for multiple imputation in this setting.

10. SUPPLEMENTS

10.1 Choice of Objective Function for Weighting

A number of objective functions have been proposed for calculating an optimal fitted table (usually in the context of contingency tables, *cf.* Fagan and Greenberg 1988). In each case the function takes the form $T = \sum T_1(W_i)$, where T_1 takes one of the forms displayed in Table 9. Each of these functions can be standardized to an equivalent function T_0 by multiplication by a constant coefficient and adding a linear function of W , so that $T_0(1) = 0$, $T'_0(1) = 0$, $T''_0(1) = 1$. Since $\sum W_i$ is constrained to a given value, the optimum weights will be unaffected. Then the standardized objective functions agree through the second term of their Taylor expansions about 1, and should give similar results when the weights are close to 1.

Table 9
Comparison of objective functions for table fitting

Name of fitting procedure	Objective function $T_I(W)$, usual form	Objective function $T_0(W)$, standardized form	Second derivative $T''_0(W)$
Least squares (minimum variance)	$(W - 1)^2$	$(W - 1)^2/2$	1
Raking	$W \log W$	$(W \log W) - W + 1$	$1/W$
Maximum likelihood	$-\log W$	$W - 1 - \log W$	$1/W^2$
Minimum χ^2	$(W - 1)^2/W$	$(W - 1)^2/2W$	$1/W^3$

in the degree of asymmetry between the costs of downweighting and upweighting, determined by the exponent of W in the second derivative, $T''_0(W) = W^{-k}$. The least squares procedure ($k = 0$) treats up-and down-weighting completely symmetrically and therefore may yield zero or negative weights. As k increases, the cost of upweighting becomes smaller relative to that of downweighting. All of the other objective functions ($k > 0$) give every observation in the raw data a positive weight; in the case of the “raking” function, this is obvious from the form of the weights as shown in Section 3.3. The use of the “raking” function here in preference to maximum likelihood or minimum χ^2 is motivated by the simple form of its solution and by the analogy to raking in contingency tables. Cressie and Read (1984) systematically study the properties of this family of measures of fit.

10.2 A Cyclic Descent Methodology for Fitting Weights

In this section we present a fitting methodology analogous to iterative proportional fitting (IPF) in contingency tables. In IPF, the cell counts are transformed multiplicatively in such a way that the cross-products are preserved (the condition for minimization of the objective function) while the table is made to conform to each set of marginal constraints in turn. The algorithm converges to a table that satisfies all of the constraints, and perforce preserves the cross-products as well (Bishop, Feinberg and Holland 1974; Ireland and Kullback 1968).

In our setting, the weights are required to have the log-linear form $W_i = \exp(a_i'\lambda - 1)$ derived in Section 3.3 while satisfying the constraints $AW = B$. In this exposition we will assume that the total weight constraint $\sum W_i = H$ is omitted from $AW = B$, and that A is of dimension p (constraints) $\times I$ (number of household compositions). We will proceed through a series of steps in each of which each weight W_i is multiplied by $c\rho^{a_{ji}}$ to obtain a new weight W'_i , thus preserving the log-linear structure; c and ρ are chosen so that the constraints $\sum W'_i = H$ and $\sum W'_i a_{ji} = b_j$ are satisfied. By proceeding cyclically so that $j = 1, 2, \dots, p$ indexes each constraint in turn, the algorithm eventually converges to weights that satisfy all of the constraints.

On step j of cycle t , the new weights are given by $W_i^{(t,j)} = c\rho^{a_{ji}}W_i^{(t,j-1)}$ (initialized for $j = 1$ by using the last weights from the last cycle, $W_i^{(t,0)} = W_i^{(t-1,p)}$). Then c and ρ must satisfy

$$\sum_i c\rho^{a_{ji}}W_i^{(t,j-1)} = H, \quad \sum_i a_{ji}c\rho^{a_{ji}}W_i^{(t,j-1)} = b_j. \tag{5}$$

Eliminating c from these equations, ρ is a root of

$$\sum_i \left(Ha_{ji} - b_j \right) W_i^{(t,j-1)} \rho^{a_{ji}} = 0. \quad (6)$$

We must have $Ha_{j,\min} \leq b_j \leq Ha_{j,\max}$ where $a_{j,\min}$ and $a_{j,\max}$ are respectively the minimum and maximum values of a_{ji} . If this were not the case, constraint j could not be satisfied with any weights. Thus there must be at least one root ρ , and if the a_{ji} are non-negative, the expression is increasing in ρ so this root is unique. The actual value of ρ is determined then by Newton's method, or by a closed-form formula for the roots of a polynomial (since with the original A , a_{ji} is the number of class j members in a household, which is an integer rarely exceeding 2).

While we have not yet proven that this algorithm always converges, we have found it to be successful in practice. This algorithm does not require any matrix inversion, and if the a_{ji} are small integers, then at each step, the recalculation of the weights involves calculating only a few integral powers. Furthermore, if some constraints take the form of simple marginals, the adjustment for those constraints takes the form of a conventional raking step.

If the original constraint matrix A is used, the procedure may take advantage of the sparseness of A (which is a consequence of the fact that only a few classes are represented in each household). At each step (say, adjusting to fit margin b_j), only the weights corresponding to non-zero a_{ji} need be modified; thus only S_1 multiplications (the number of nonzero entries in A , which is less than the population of the block) and perhaps $3S_1$ additions are required per cycle, as compared to $5S_1 + S_2$ operations per iteration with Newton's method. On the other hand, the rows of A tend to be highly dependent, so convergence may be slow (typically 20 cycles in our simulations); orthogonalization of A destroys the sparse structure of the coefficients. Thus, unless S_2 is much larger than S_1 (or unless some other method is devised to accelerate the algorithm), raking is not faster than Newton's method.

ACKNOWLEDGEMENTS

This research was supported by Joint Statistical Agreements 86-8 and 87-7 between the U.S. Bureau of the Census and Harvard University. The author has benefited from comments by Donald B. Rubin and other participants in the seminar on census undercount at the Harvard University Department of Statistics, and Nathaniel Schencker, Nash Monsour, and other members of the Undercount Research Staff in the U.S. Bureau of the Census. Simulations used data made available by the Inter-university Consortium for Political and Social Research and originally collected by the U.S. Department of Commerce, Bureau of the Census.

REFERENCES

- ALEXANDER, C.H. (1987). A class of methods for using person controls in household weighting. *Survey Methodology*, 13, 183-198.
- BISHOP, Y.M.M., FIENBERG, S.E., and HOLLAND, P.W. (1974). *Discrete Multivariate Analysis*. Cambridge: M.I.T. Press.
- BUREAU OF THE CENSUS (1985). Census of Population and Housing, 1980: Public Use Microdata Samples.

- CILKE, J.M., and WYSCARVER, R.A. (1988). The Individual Income Tax Simulation Model, in Office of Tax Analysis, *Compendium of Tax Research* 1987, Washington: Government Printing Office.
- COX, L. (1987). A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association*, 82, 520-524.
- COX, L., FAGAN, J., GREENBURG, B., and HEMMIG, R. (1986). Research at the Census Bureau into disclosure avoidance techniques for tabular data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 388-393.
- CRESSIE, N., and READ, T.R.C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society, Series B*, 46, 440-464.
- DEMING, W.E., and STEPHAN, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11, 427-444.
- DIFFENDAL, G. (1988). The 1986 Test of Adjustment Related Operations in Central Los Angeles county. *Survey Methodology*, 14, 71-86.
- FAY, R.E. (1986). Implications of the 1980 PEP for future census coverage evaluation. U.S. Bureau of the Census, unpublished.
- FAGAN, J.T., and GREENBERG, B. (1988). Algorithms for making tables additive: Raking, Maximum Likelihood, and Minimum Chi-square. *Proceedings of the Section on Survey Research Methods, American Statistical Association* (forthcoming).
- GASS, S.I. (1964). *Linear Programming: Methods and Applications*. New York: McGraw-Hill.
- IRELAND, C.T., and KULLBACK, S. (1968). Contingency tables with given marginals. *Biometrika*, 55, 179-188.
- LITTLE, R.A., and RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.
- NATIONAL ACADEMY OF SCIENCES (1985). *The Bicentennial Census: New Directions for Methodology in 1990*. Washington: National Academy Press.
- OH, H.L., and SCHEUREN, F.J. (1978). Multivariate ratio raking estimation in the 1973 Exact Match Study. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 716-722.
- PURCELL, N.J. (1979). Efficient estimation for small domains: a categorical data analysis approach. Ph. D. dissertation, University of Michigan.
- PURCELL, N.J., and KISH, L. (1979). Estimation for small domains. *Biometrics*, 35:365-384.
- RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.
- SCHEUREN, F.J. (1981). Methods of estimation for the 1973 exact match study. In *Studies from Interagency Data Linkages*, Washington: Social Security Administration.
- ZASLAVSKY, A.M. (1989). Representing Census undercount at the household level. Ph. D. thesis, Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts.

QUID, A General Automatic Coding Method

JACQUES LORIGNY¹

ABSTRACT

The QUID system, which was designed and developed by INSEE (Paris) Institut National de la Statistique et des Études Économiques – National Statistics and Economic Studies Institute, is an automatic coding system for survey data collected in the form of literal headings expressed in the terminology of the respondent. The system hinges on the use of a very wide knowledge base made up of real phrases coded by experts. This study deals primarily with the preliminary automatic standardization processing of the phrases, and then with the algorithm used to organize the phrase base into an optimized tree pattern. A sorting example is provided in the form of an illustration. At present, the processing of additional coding variables used to complement the information contained in the phrases presents certain difficulties, and these will be examined in detail. The QUID 2 project, an updated version of the system, will be discussed briefly.

KEY WORDS: Automatic coding; Natural language variables; Phrase matching; N-grams.

1. INTRODUCTION

The QUID (abbreviation of QUEstionnaires d'IDentification – Identification Questionnaires) system is an automatic coding system designed and developed by the Institut National de la Statistique et des Études Économiques (INSEE – National Statistics and Economic Studies Institute) in 1979-1980.

Review of the Problem

The problem consists of automatically classifying an individual surveyed into a job defined in accordance with an existing nomenclature (for example, the nomenclature of the professions). In order to do this, the system uses mainly the natural language answer given in response to a direct question (for example, “What is your present profession or trade?”), as well as additional information contained in the survey form, which is assumed to have been previously coded (for example, the Economic Activity code for the firm where the individual works).

In our terminology, a direct answer in natural language is called the “literal heading”, or simply “heading”. Any additional encoded information is represented by the generic term “additional variables”.

In the next section, we will discuss the basic approach of the QUID system and the results of its implementation at INSEE. In section 3, we will describe the present version of the system. Finally, in section 4, we will examine the problems surrounding the processing of additional variables, and will discuss the new version of the system (QUID 2), which should help resolve the difficulties encountered.

¹ Jacques Lorigny, Administrateur à l'Institut National de la Statistique et des Études Économiques 18, Bld Adolphe Pinard 75675 PARIS CEDEX 14 (France).

2. THE PRINCIPLE BEHIND THE METHOD

2.1 The Basic Approach

The basic approach of the QUID system consists of building a very large data base made up of typical respondent headings accompanied by a corresponding code assigned by an expert. The data base is as large as possible in order to make it possible to obtain a high matching rate, and new headings are added to the base as they appear.

In our terminology, the data base is called a "knowledge base" or "knowledge file" (KF), because it has the ordinary structure of a flat file in its raw state. Most often, the knowledge file is set up on the basis of a survey carried out during a previous year, which has already been coded either manually or using an interactive method. Each base heading is accompanied by its code (which is *a priori* assumed to be accurate), and its "frequency of occurrence" in the KF; that is, the number of individuals who responded using this heading.

The management task of the knowledge base (auditing, expansion) is completely separate from the operation of coding the survey under way. It is the responsibility of a central office staffed by expert coders, while the coding operation itself is most often regionally decentralized.

The difficulties of an approach of this type derive from the rapid increase in the time required to search the base as it grows in size. In order to solve this problem, the QUID system uses mathematical results derived from Information Theory (Shannon 1948; C.-F. Picard 1972; B. Bouchon-Meunier 1978; M. Terrenoire 1970; D. Tounissoux 1980), which can be used to minimize search time by organizing the base in the form of an optimized tree structure.

The basic approach of the QUID system also makes it possible to opt for a set of general programs; that is, those that can be used with all semantic fields, for example, professional, food products, or municipal headings.

2.2 Results

The system has been tried for various INSEE tasks and is presently being used to code the CS (socio-professional category) code in order to process DADS (Déclarations annuelles de données sociales - Annual social information) data provided by all firms that employ paid labour. The following figures provide an idea of the orders of magnitude involved.

At present the knowledge file for the DADS application contains 122,000 headings (representing a knowledge base population of 650,000 wage earners). Its optimized organization consists of a tree with about 100,000 nodes (of which 86,000 represent certainty nodes; see section 3.2). It has been used to code a population of 570,000 wage earners with an average effectiveness of 90%, varying between 85% and 95%, depending upon the region. By "effectiveness", we mean the percentage of cases where the system provides a single answer which is accepted on principle under the conditions of this application. At present, since we do not have a precise measurement of the validity of these single answers, we estimate that the error rate is likely to be in the order of 5% to 10%. However, the knowledge base is being audited by the Dijon Expert Centre, according to which a significant proportion of the error rate should normally decrease. Once this has been achieved, we will have more accurate figures to report.

From the point of view of data processing limitations, the optimized tree is loaded into 3,300 kilobytes of central (virtual) memory and the automatic coding time for an individual case is in the order of 40 ms in an IBM 4341 central processing unit.

For the last few months, we have had available a variant of the coding program itself. This has been designed for use with mini-computers and can load the tree by sections, depending upon available memory space.

In applications other than DADS data, effectiveness is not as high, no more than 75%. It all depends upon the quality and comprehensiveness of the knowledge base.

3. THE PRESENT VERSION OF THE QUID SYSTEM (QUID 1)

3.1 Preliminary Standardization of Headings

Before constructing the optimized tree, the raw headings are first standardized in accordance with a set of external parameters chosen by the user for his application.

The words are separated and fitted into predetermined zones whose length (a single one for all words) and maximum number (a single one for all headings) are parametrized. It is advisable to choose a larger value on the basis of these two parameters, and allow the optimization algorithm itself to select the significant elements of the heading (see section 3.2). For example, the DADS application (see section 2.2) chose 4 zones of 12 characters each.

“Empty words” are eliminated. The list of empty words is an external parameter provided by the user for his application. Most often, it includes articles, prepositions, *etc.*, and is significantly dependent upon the application.

Initials are standardized (I.N.S.E.E. becomes INSEE, S N C F becomes SNCF).

Finally, the user may process the table of separate words in any way he wants (in the form of a subprogram in the PL/1 language). In fact, this is rarely necessary and seldom used (except to code municipal codes from municipal headings).

Once word processing has been completed, the words are divided into bigrams (blocks of two consecutive letters) or trigrams (blocks of three consecutive letters), *etc.* Choosing the type of blocking is parametrized (however, a single parameter is used for the entire application). In practice, blocking into bigrams is the only type that has been used until now; however, the idea of blocking into trigrams should be tested. For the purposes of this study, we will only consider blocking into bigrams.

3.2 The Algorithm Used to Set Up the Optimized Tree Pattern

Let $T = (t_1, t_2, \dots, t_j, \dots, t_n)$ the code to be coded, for example all the modalities of the Profession code.

$Q = (q_1, q_2, \dots, q_i, \dots, q_m)$ all the bigrams resulting from the standardization of the headings (for example, $m = 24$ when the number 4 has been chosen as the “number of words” parameter, and 12 characters as the “word length” parameter).

X = the tree pattern to be constructed, which we call a “QUID” (questionnaire d’identification – identification questionnaire).

The algorithm constructs X by parsing down from the root node x_0 (which by convention is placed at “level 0”) to the nodes in levels 1, 2, *etc.*

At root node x_0 it links the entire KF, and searches for the best bigram to query first; that is, that which can discriminate best for the desired code T in the entire KF.

$N(x_0)$ represents the total frequency of occurrence associated with the entire KF; that is, the sum of frequencies accompanying the base headings,

$N(x_0, j)$ is the frequency of occurrence of code t_j in the entire KF.

We assume that the knowledge population is statistically representative of the population to be coded (we should recall that, in practice, the KF is often the survey file for a previous year).

Thus, we can estimate the probability of finding code t_j the population to be coded on the basis of the following formula:

$$\Pr(t_j | x_0) = N(x_{0,j})/N(x_0).$$

The *a priori* ambiguity for T is measured on the basis of Shannon's entropy:

$$H(T/x_0) = \sum_j \Pr(t_j | x_0) \log 1/\Pr(t_j | x_0).$$

Let us assume that a bigram, for example q_i is allocated to node x_0 . To each of its modalities in the KF, we associate the sub-base made up of the headings that have this modality.

Let $(a_i^1, a_i^2, \dots, a_i^k, \dots)$ represent the modalities captured by bigram q_i in the KF. For each of these modalities, thus, for each of the sub-bases generated, we create a node y , which follows immediately after x and is located at level 1 of the tree.

The information provided by bigram q_i (which is assumed to be assigned to root node x_0) is measured by the average reduction in the ambiguity of T when we go from x_0 to one of the y nodes.

That is:

$$I(x_0, T, q_i) = H(T | x_0) - \sum_{y \in \Gamma(x_0)} \Pr(y) H(T | y),$$

where

$\Gamma(x_0)$ represents all the successive y nodes at level 1 below node x_0

$H(T | y)$ the conditional entropy of T at node y .

(same formula as above but replacing x_0 by y).

$\Pr(y) = N(x_0, a_i^k)/N(x_0)$ if a_i^k is the modality of bigram q_i which generates node y , and $N(x_0, a_i^k)$ is the frequency of occurrence of modality a_i^k of bigram q_i in the entire KF.

The algorithm carries out this data calculation for all bigrams q_1, q_2, \dots, q_m , because at root node x_0 they are all possible candidates for selection as the first bigram to be queried.

The algorithm chooses the bigram which maximizes $I(x_0, T, q_i)$. For example, in the case of q_{i0} , it effectively divides the base into as many sub-bases as there are modalities of bigram q_{i0} in the base. This effectively creates the y nodes that follow x_0 at level 1, and the construction of level 1 of X is thus completed.

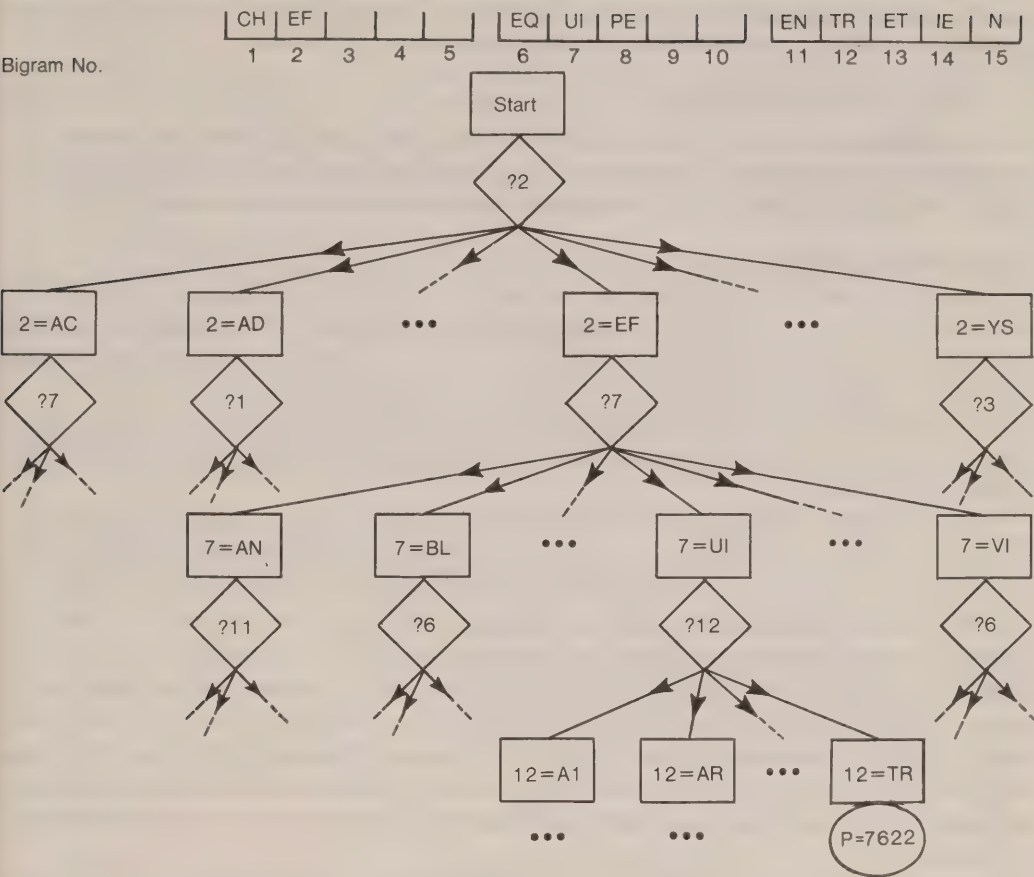
For each sub-base obtained (thus, for each y node), the algorithm carries out exactly the same operation as that which we have just described for root node x_0 , and so on.

The process stops for a given node:

- (1) when there is only one heading at the node; in this case, the conditional entropy is zero; or
- (2) when there is only a restricted number of headings that differ in terms of the remaining bigrams, but which have all the same code; or
- (3) when there are two headings or more, but they have different and not distinguishable codes.

Cases (1) and (2) are known as "certainty nodes", and case (3) is known as the "uncertainty node". Together, they represent the "terminal nodes".

Standardized Heading:



- Query the content of bigram no. 2 in this node of the tree.
- The content of bigram no. 2 is EF.
- In this node of the tree, we may determine the profession code: its value is 7622 (1975 Trade Nomenclature).

In this example, the raw heading is that of the profession entered by the individual surveyed. The objective of the system is to find the corresponding profession code in the 1975 Trade Nomenclature.

Initially, we extract the first ten characters of the three most significant words. In this way, we obtain the standardized heading, which is then blocked into pairs of letters (these are called bigrams and are numbered from 1 to 15). Then, we query the system. This operates in accordance with a chain of questions and answers optimized by a mathematical algorithm based on information theory. This calculation takes place during the course of a preliminary phase which determines the first bigram queried as a function of a given knowledge file, and then the following sequence of questions depending upon the answer obtained each time. At this point, the computer queries first bigram no. 12, which contains TR. At this stage, it ascertains that it can without ambiguity determine that this represents the Profession 7622 code (Technical Staff and Technicians). On the average, processing time takes a total of 41 milliseconds of computer time in an IBM 370/148, and the amount of central memory used is 380 Kbytes.

Raw Heading: head, maintenance team.

Figure 1. Example of Classification of a Heading in the Tree.

The construction of the tree X continues from level to level until the KF has been exhausted. In fact, we have never gone beyond level 15, but there is no set limit for the system itself. An example of classification in the tree is shown in Figure 1.

3.3 The Use of the Coding Itself

In order to code a heading in the current survey, we start by standardizing the information in accordance with section 3.1. Then, the bigrams obtained are matched against those of the Quid loaded into the computer. The exploration leads to three possible results.

3.3.1 Certainty Node

The system provides a single code but this may well be wrong if the knowledge base is not comprehensive enough. For example, during one of our first tests in 1979, we obtained a certainty node for level 1 on the basis of bigram $2 = CC$, since the only heading obtained had been VACCINEUR VOLAILLES (POULTRY VACCINATOR).

When we later had to code the heading RACCOMMODEUR VÊTEMENTS (GARMENT MENDER) the single code obtained was that representing agricultural service professions, and the error was obvious.

Thus, we added to the system a control procedure based on single echoes. This process is known as "redundancy control" and consists of verifying, after the detection of a single echo, the content of the first three bigrams of each word. A single echo (obtained on the basis of the vector leading to a certainty node) is said to be non-ambiguous, when the cluster of headings in the certainty node contains at least one heading that has the same redundancy bigrams as those of the heading to be coded. Otherwise, the echo is said to be ambiguous, and consequently treated as an anomaly of the automatic system. Experience has shown that this arrangement tends to consolidate significantly the reliability of the system without appreciably overburdening the tables in memory or increasing processing time (even in large applications, the number of redundancy formulas per certainty node is, on the average, in the order of one, and rarely goes beyond ten).

In order to be thorough, we should add that this redundancy control is not rigidly set once and for all. The user has two external parameters: the list of bigrams over which he intends to exercise control, and the (maximum) number of bigrams retained. In this way, he can keep in check the severity of the matching control, depending upon his objectives in terms of the quality and "effectiveness" of automatic coding.

3.3.2 The Uncertainty Node

The system provides various possible codes (most often two codes), and displays their respective frequencies of occurrence at the node under consideration. In this case, the officer who has the file of the survey being processed will then manually reject one of the two.

3.3.3 The Case of an Unknown Response

If, during the course of exploring the Quid, the modality sought is not found in the modalities captured by the bigram queried, the search will fail and this also represents a case of rejection that must be processed manually.

New cases encountered during the course of processing will be stored in memory, centralized in the expert centre, verified, and then incorporated into the KF in order to produce a new expanded version of the Quid.

At present, for purposes of convenience, the knowledge iteration takes place once a year, but nothing prevents it from being organized so that it takes place more often so that applications can progress faster, for example in the case of population surveys.

4. THE PROBLEM OF PROCESSING ADDITIONAL VARIABLES

In the present version, QUID 1, additional variables are simply structured into bigrams and processed in the same way as literal data. This leads to certain difficulties and problems that made it necessary to develop a new version, QUID 2, which operates in two stages:

- in the first stage, QUID 1, which is reserved for processing the literal heading and producing either the final code (when this is totally determined by the heading), or an internal code designating a rule or decision table that can be applied to the additional variables to achieve the calculation;
- in the second stage, the rules or decision tables achieve the determination of the final code.

Detailed Examination of the Difficulties Encountered

At times, certain nomenclatures that are particularly complex, such as the PCS Code (Nomenclature of Professions and Socio-Professional Categories) call upon a combination of the literal heading and various additional variables.

For example, the coding of the PCS code uses the Professional Category additional variable (which is abbreviated to CPF). The following is the question such as it appears in the 1982 Population Census Individual Form:

- Indicate the professional category of your present job:
- | | | |
|----------------------------------|---------------------------------------------------|---|
| | - unskilled or semi-skilled labourer | 1 |
| - labourer | - semi-skilled labourer (OS, O1, O2, O3, . . .) | 2 |
| | - skilled labourer (P1, P2, P3, TA, OP, OQ . . .) | 3 |
| - clerk | | 4 |
| - technician, draftsman | | 5 |
| | - supervising workers or clerks | 6 |
| - foreman | - supervising other foremen or technicians | 7 |
| - engineer or professional staff | | 8 |

The additional question was made necessary by the fact that the heading alone is not always enough to classify the individual in accordance with PCS nomenclature.

For example, a LUMBER COMPANY WORKER

- must be classified into 6916 (lumber company or forestry worker) if his CPF is 1, 2, 3, or 4
- and into 4801 (Managerial and supervisory staff of agricultural or lumber operations) if his CPF is 5, 6, 7, or 8.

The present system considers these additional variables as if they were literal data. They are placed at the end of the heading and structured into bigrams in the same way (for example, the CPF variable with the addition of a blank space is placed into the (m + 1)th bigram). However, this solution is not satisfactory and leads to various errors:

Error No. 1. When there is not enough information in the KF, this may lead to many cases of unknown responses.

For example, if the KF has only one LUMBER COMPANY EMPLOYEE with a CPF = 2 and another with a CPF = 7 the file will be unable to find a LUMBER COMPANY WORKER

with a CPF other than 2 or 7 (that is, *a priori* in 6 cases out of 8). This error is made worse when the additional variable is very diluted, for example, in the case of the variable representing the Economic Activity of the undertaking (which is abbreviated as additional variable AE).

Error No. 2. When there is not enough information in the KF, this may lead to miscodings. For example, if the KF has only one LUMBER COMPANY WORKER with a CPF = 2, the CPF bigram will not discriminate or appear in the search key, so that a LUMBER COMPANY WORKER with a CPF = 7, will be classified into PCS = 6916 instead of 4801. This is a case of miscoding

In order to correct this defect in the present system, the only measure we can take is to apply the redundancy control to the additional variables (and thus obtain an ambiguous or questionable case which is rejected or corrected manually, instead of allowing the error to remain undetected). However, here again, this is only a last resort. In fact, the additional variables lead to an unchecked expansion of the KF. Each KF reference has its own cross combination of modalities of additional variables, and it is not very likely that we would find the same combination for a new individual to be coded. Thus, this will lead to many uncertain cases and automatic coding rejections, which will reduce the practical benefits of mass exploitation.

The two errors, no. 1 and no. 2, are related to the relative incompleteness of the KF. For example, it would be enough to enter into the KF eight LUMBER COMPANY WORKER titles and add in each case one of the possible CPF modalities (1 to 8), in order for the two errors to disappear. However, in the case of real applications, we find that the relative incompleteness of the KF decreases quite slowly, as it grows to reach its operating pace. Contrary to the lexicographic space of literal headings, which tend to become dense rather quickly, the cross checked space of the additional variables remains a vast frontier for a long time, and goes very slowly from a density of occupation of 0 to a density of 1 (one individual).

Error No. 3. There is a third category of errors that are not caused by the incompleteness of the KF but by the excessive sensitivity of the QUID in relation to errors inevitably contained in the file (and this always in relation to the additional variables).

Let us take a simple example. Let us assume that the SENIOR SECRETARY heading must be coded PCS = 4615 (senior secretarial staff), regardless of the value of all the additional variables. Let us consider the following KF, in which an error has slipped by (for example, the failure to assign the PCS code):

Heading	CPF a.v.	AE a.v.	PCS Code
Senior Secretary	[7]	[49 11] (fashion design, haute couture)	4615
Senior Secretary	[7]	[83 43] (loan cooperative)	4616 ↑ error

Even though the AE additional variable should not be used to code the PCS code, the QUID algorithm uses it to separate the two certainty nodes.

- One in favour of 4615 in view of bigram AE1 = 49.
- And the other in favour of 4616, in view of bigram AE1 = 83.

The result is that, during the coding stage itself, all senior secretaries belonging to economic sectors other than those starting by 49 or 83 will appear as “unknown cases”. Moreover, those in all sectors starting by 83 will obviously produce errors. However, it is mainly the first phenomenon that interferes with accuracy, because it affects an area that is much larger than that affected by the initial error.

Error No. 4. Finally, the present QUID algorithm is excessively rigid in terms of choosing the optimal question. Most often, this results in a simple inversion of the order of the questions in the course of the search, in relation to the order that would have been preferred by the designer. Thus, the effect is secondary, since the final results are identical. However, this may also lead to more serious distortions.

Let us take the following (partly fictitious) example. Let us assume that, according to the nomenclature, the SENIOR SECRETARY heading should be coded either PCS = 4615 as above if the CPF additional variable CPF = 1 to 7, and PCS = 3726 (current managerial staff in other administrative business services), if CPF equals 8.

Let us examine the KF containing the following two references:

Heading	AE a.v.	CPF a.v.	PCS Code
Senior Secretary	<u>49</u> <u>11</u>	<u>8</u>	3726
Senior Secretary	<u>83</u> <u>43</u>	<u>7</u>	4615

Thus, the two references are correctly coded. When the QUID algorithm arrives at a node where it has examined all the possible bigrams of the literal heading, it must now choose one bigram in the additional variables, in order to separate the two final results: PCS = 3726 and PCS = 4615. In this simple but not altogether unrealistic example, the three possible bigrams: AE1, AE2, and CPF, provide the same quantity of information (one bit). In our algorithm, the arbitrary convention is that in cases of equality, the program should choose the first question in the order in which the additional variables were presented in the form. However, in this example, this will be deceiving, since we would encounter the aberration discussed above (error no. 3). However, it is not possible to determine an order of additional variables that would prevent this type of error in all important cases. We can only seek an order of questions that will be statistically the least invalid, by groping our way on the basis of the order of conceptual splits, the negentropic capacity of each additional variable, *etc.*

5. CONCLUSION

In its QUID 1 version, the present QUID system provides very valuable services to INSEE. Nevertheless, it still has certain weak points regarding the processing of additional variables. The new QUID 2 version should improve processing while remaining faithful to our “basic approach” to the automatic coding problem, which could be summarized in two points:

1. Separation of the knowledge base (in this case, a base of rules and decision tables that are written in natural language, are independent of each other, and are audited and managed by an autonomous expert centre), and the use of automatic coding programs (in this case, loading and table exploration programs).
2. Construction of general programs; that is, programs that are independent of the semantic field processed.

At least, these are the objectives that we try to attain.

ACKNOWLEDGEMENTS

I would like to thank the reviewers for their invaluable help during the preparation of this paper.

REFERENCES

- BOUCHON-MEUNIER, B. (1978). Sur la réalisation de questionnaires. Doctoral thesis. Paris.
- KNAUS, R. (1987). Methods and problems in coding natural language survey data, *Journal of Official Statistics*, 3, 45-67.
- LORIGNY, J. (1982). Mesures d'entropie et d'information pour les systèmes ouverts complexes. Doctoral thesis, Paris.
- LORIGNY, J. (1985). Manuel d'utilisation du système QUID. Institut National de la Statistique et des Études Économiques, Direction de la production, Paris.
- PICARD, C.-F. (1972). *Graphes et Questionnaires*. Paris: Gauthier-Villars.
- SHANNON, C.E. (1948). A Mathematical Theory of Communication. *Bell Systems Technical Journal*, 27, 379-423, 623-656.
- TERRENOIRE, M. (1970). Un modèle mathématique de processus d'interrogation: les pseudo-questionnaires. Doctoral Thesis, Grenoble.
- TOUNISSOUX, D. (1980). Processus séquentiels adaptifs de reconnaissance de formes pour l'aide au diagnostic. Doctoral Thesis, Lyon.

ACTR

A Generalized Automated Coding System

M.J. WENZOWSKI¹

ABSTRACT

A generalized implementation of a method for performing automated coding is described. Traditionally, coding has been performed manually by specially trained personnel, but recently computerized systems have appeared which either eliminate or substantially reduce the need for manual coding. Typically, such systems are limited in use to those applications for which they were originally designed. The system presented here may be used by any application to perform coding of English or French text using any classification scheme.

KEY WORDS: Automated coding; Classification; Text searching.

1. INTRODUCTION

Automated coding refers to the process by which text is machine analysed in order to assign it a classification, or code. To be practical, automated coding systems must be capable of coping with such problems as: rearranged words, plural vs singular forms, missing words, extraneous words, spelling variations, synonyms, abbreviations, inconsistent hyphenation and variable punctuation and syntax. In addition, in searching a text database for a match, they should be capable of determining the closest match when no identical match can be found.

Generalized systems provide all of the features required, packaged within an easy to use, flexible, and efficient framework. To use a generalized system for a particular application, no development or conversion effort is needed to tailor it to the application specific requirements. As well, no application sponsored support for the maintenance of a generalized system is necessary, since the package is supported and maintained by a central agency.

ACTR (an acronym for: Automated Coding by Text Recognition) employs techniques similar to those employed in other automated coding systems currently in production at Statistics Canada (Landry and Pidcock 1984), but is unique in that it has been generalized to allow it to be used by any application to assign codes based on the input of English or French text according to any classification scheme.

The methods which ACTR uses to perform automated coding are based on techniques which were originally developed at the U.S. Bureau of the Census (Appel and Hellerman 1983). Basically stated, the method consists of searching through a collection of text previously associated with correct codes. If the subject text is successfully located, the associated code is returned and the process ends. Otherwise, the search continues, but uses an algorithm to locate the closest match, and subsequently assign its associated code.

¹ M.J. Wenzowski, Research and General Systems, Statistics Canada, Room 2306, Main Building, Ottawa, Ontario, K1A 0T6.

2. USING ACTR

To use ACTR in an automated coding application users first need to define the text and associated codes which they intend to use as a standard for matching. While there are many sources for this information, the best is a set of text which is representative of the text which will most likely be encountered in a matching run. For a survey, this generally means the responses and manually assigned codes from a previously completed survey. Although great care should be taken to ensure that the correct codes have been assigned, the text should be left as is, complete with spelling, grammar and syntax errors, since in this form it is most representative of the text which will be encountered in subsequent surveys.

After having defined a file of text and correctly assigned codes, they must be loaded into a matching database. ACTR provides the software required to perform this task and so automatically transforms the file into a matching database.

ACTR has been designed to allow an iterative approach to developing an automated coding application. Accordingly, text and codes can be added, changed or deleted at any time during the life of the application. In addition, the parsing strategy (discussed in detail below) can be altered at any time. Thus, users are presented with a software framework which, through cycles of database updates and matching runs, will allow for as many iterations as is necessary to obtain the matching quality desired. Users are encouraged to use ACTR in this manner, since ultimately it leads to higher quality and more economical coding operations.

3. PRINCIPLES OF OPERATION

In the case of a human being performing a coding operation, the similarity between occupations described as "Computer Programmer" and "Programming Computers" is so great that they would generally be judged as identical. However intuitive this reasoning may seem, computer systems in general would rate the two as unequal. Unfortunately, natural language (for example, English or French) frequently provides a large number of ways to express the same meaning. So, for a computer based system to be able to cope with this variance, there must be some means by which a degree of similarity can be determined.

This is the essence of ACTR: text is rated according to how similar it is to some other text. In the preceding example, ACTR treats the two occupation descriptions as identical since, after suffixes are truncated, double letters are removed and word order is ignored, both phrases become "Comput Program" and as such are clearly equal.

The steps employed in reducing the above phrases to a standard form are part of what is known in ACTR as the parsing strategy. ACTR's parsing strategy is entirely user controlled and may be changed at any time during the life of an application. Users exercise control over the parsing strategy employed in their applications by supplying the data which is to be used to direct the process. This means that all steps are entirely controlled by the user, even to the extent of allowing a step to be skipped.

The Parsing Strategy

Parsing is the ACTR process which is responsible for the reduction of phrases to a standard form. Ideally, the resulting form should be such that any two phrases with the same words will be identical in their ACTR representation regardless of their syntactical and grammatical differences. Returning to the previous example, the two phrases "Computer Programmer" and "Programming Computers" when properly parsed, should ideally result in a set of identical words for each phrase. For example, both phrases could be reduced to "Comput Program".

The parsing process employed may involve the reduction of plural forms, elimination of trivial words, removal of suffixes and/or a number of other steps. Although the order of the parsing steps applied is fixed by ACTR, users control how, if at all, each step is executed. For further information on the order of parsing, the interested reader should consult Connor, Salloum and Wenzowski (1988).

Basically, the parsing process can be thought of as having the following two major subcomponents:

1. **TEXT PROCESSING.** In this stage of parsing, the text supplied is processed as a continuous stream of characters. Although one may think of the text as containing words, spaces and punctuation, none of these is given any special consideration at this point in the parse. This view is necessary in order to allow for the recognition of particular character strings exactly as they occur *in situ*.
2. **WORD PROCESSING.** When this stage of the parse begins, the text has already been broken down into words and so further processing is performed on a word by word basis. This view is necessary since a large amount of text standardization occurs on the basis of defined words.

Text Processing

As already discussed, these steps are performed regardless of context. Thus, the following steps are performed on a character by character basis.

Exclusion Clauses: Exclusion clauses are ignored in matching, but are used in database updating to indicate the intention of allowing controlled duplication of phrases. By default, ACTR will not allow identical phrases to be loaded into a matching database.

By providing a means of controlling duplication, users are able to load phrases which could have more than one code assigned, even though they are identical after having been parsed. Although not used in matching, exclusion clauses are stored along with the phrase in the matching database and can subsequently be used to manually resolve multiple matches.

The syntax of an exclusion clause is defined entirely by the user. Both beginning and terminating strings must be provided. These and any information enclosed by them are ignored during matching.

As an example, consider an exclusion clause syntax defined with a beginning string of "(Except" and a terminating string of ")". With this in place, the two phrases "Computer Programming (Except As An Employee)" and "Computer Programming (Except As Self-Employed)" could co-exist in the matching database, even though their ACTR representations are identical. Subsequently, if a match for "Computer Programmer" is requested, both of these phrases would be returned. Since exclusion clauses are stored along with the original phrase text, they can be displayed to a reviewer, who could then manually resolve the match.

Deletion Strings: If any deletion string supplied by the user is found in any position in a phrase, ACTR will remove it from consideration before continuing the parse.

As an example, in English processing, this is a way in which the apostrophe can be removed. For example, the two phrases "Electrician's Apprentice" and "Apprentice Electrician" would become identical with the removal of the apostrophe.

Note that if this step were not performed, the apostrophe would most likely be used as a word delimiter. This would yield three words for the first phrase and two for the second, of which only one word would be common to both.

Replacement Strings: This facility is most useful for standardizing abbreviations. This is desirable since abbreviations commonly include characters which, although useful to the abbreviation, would be viewed as word separators at a later stage in the parse. If this were allowed to happen, information loss would most likely occur.

As an example, if the string "T.V." was defined with a replacement value of: "Television" then any occurrence of the original string would be translated to the replacement value before continuing the parse.

Note that if this step were not performed, the result of parsing "T.V." would most likely be the two letters "T" and "V". This is clearly undesirable, since the meaning of the abbreviation has been completely lost.

Word Characters: ACTR defines a word as any contiguous sequence of characters in a phrase which are all members of the set of characters contained in the word character list. Any characters not in this list will be used as word delimiters and will be dropped from further consideration.

Typically, the set of word characters used contains all of the letters of the alphabet and all of the numeric characters. With this in place, a phrase of "Farmer/Fisherman" will result in two words, since "/" is not a word character and is therefore used as a word delimiter.

Word Processing

At this point, ACTR begins to treat the text as a collection of words. Thus, the following processing steps are applied on a word by word basis.

Hyphenated Words: Any hyphenated words supplied are replaced by the substitute word(s) also provided. This feature is very useful in providing for the recognition of words and word groups which are inconsistently hyphenated.

As an example, if the user defines "Take-Out" as a hyphenated word with a substitute word of "Takeout" then this substitution will be made. If, on the other hand, this definition had not been made, then two words would result if the hyphen was not a word character.

Illegal Word Characters: If any of the strings supplied are found to exist in any word in any position, then that entire word is removed from further consideration.

As an example, some applications use this feature to eliminate words which contain numeric characters. So, if the set of numeric digits was given as illegal word characters, then a word like "DEPT716A" would be removed from further consideration.

Replacement Words: This feature provides a synonym capability in order to ensure that two dissimilar words will be recognized for matching purposes. This can also be useful to overcome commonly occurring spelling mistakes.

As an example, if the phrases "Automobile Repairs" and "Car Repairs" were processed with the word "Car" given as a replacement word for "Automobile" then the two phrases would be made identical.

Double Words: This feature forces ACTR to consider not only the occurrence of the two word grouping, but their order as well. This can be useful to overcome inconsistencies in word spellings and also to preserve word order.

As an example, consider the phrase "Take Out Restaurant". Although this would yield three perfectly acceptable words, the words "Take" and "Out" would not match to either of "Takeout" or "Take-Out". However, if a double word combination of "Take Out" was defined with a replacement of "Takeout" then the first case in the example given is addressed.

We are presented here with an example of how steps in the parsing strategy can be used together. If the hyphenated word example given above was also entered, then all of the hyphenated, double word, and single word cases would match.

Trivial Words: If any word in this set is encountered in the course of parsing, then it will be removed from further consideration.

As an example, if the set of trivial words contained "A", "Am" and "I", and the two phrases "I Am A Computer Programmer" and "Computer Programmer" were encountered, then the phrases would match.

Suffixes: At this point, words are scanned right to left looking for the longest defined suffix such that the remaining word, after the suffix is removed, will be at least five characters in length. If a defined suffix is found, it is removed.

As an example, if the suffixes "ing" and "er" are defined, then the phrases "Computer Programming" and "Computer Programmer" will match.

Replacement Suffixes: Replacement suffixes are searched for in a word by scanning right to left for the presence of the longest defined replacement suffix. If one is found, it is removed and the substitute supplied is used in its place.

As an example, the user may wish a plural form to be reduced to a singular one so that the singular suffix will be recognized in the suffix truncation step. This is demonstrated with the phrases "Battery Manufacturing" and "Manufacturing Batteries". If the suffix "ies" is changed to "y" then not only will the phrases be the same, they will be processed in the same manner at suffix truncation time.

Double Letters: At this stage in the parse, each word is examined for the presence of any double character occurrences which are contained in the (user-defined) double letter set. If any are found, they are reduced to a single occurrence.

Typically, the double letter set used is the full set of alphabetic characters. If this is the case, then the words "Programer" and "Programmer" would match, in spite of the spelling error.

Root Words: At this point, words are scanned for the presence of any of the root words supplied. The scan is applied from left to right in the word, and searches for the longest defined matching root word. If one is found, then its substitute is used as a replacement for the word and the suffix truncation and replacement steps are skipped.

As an example, the languages "Slavee" and "Slavic" differ only in their last two characters. So, if the suffixes defined include "ee" and "ic" then an information loss occurs, since both words will become identical. Although generally, suffix truncation works well for most applications, it quite clearly fails for this particular example. To overcome this problem, if root words of "Slave" and "Slavi" are defined, then the suffix truncation step is bypassed for these cases only. Thus, as suffix truncation problem cases are identified, root words and their substitutes can be defined to overcome them.

Duplicate Words: Finally, the set of words resulting from the parse of the supplied text is examined for the presence of duplicates.

Note that words which are duplicates at this point may not have appeared as duplicates before the text was parsed. Only one occurrence of each word defined at this point in the parse is kept.

4. SEARCHING AND MATCHING METHODS

ACTR always processes the supplied text according to the parsing strategy defined before attempting a match. If after doing this, ACTR is able to locate a phrase on the matching database with all of its words in common with all of the words in the supplied text, then the match found is referred to as a "Direct Match". If a direct match cannot be found, ACTR may, as a user option, continue to search the database for the closest match. This latter type of match is called an "Indirect Match". Although they share a common foundation in that they are both based on parsed text, the two matching methods used by ACTR differ greatly in their mechanisms for both locating and assigning a match.

Direct Matching

In direct matching, only a 100% match is searched for. Recall that matching is based on parsed text, so phrases which are 100% matches may not appear to be identical in their original form. This is a direct effect of the parsing strategy in use.

In terms of database access techniques, the fastest path to an item is through the use of a key. Unfortunately, the roadblocks to keyed access of ACTR phrases exactly as they occur include a maximum phrase length of 200 characters and an upper limit of 20 on the number of parsed words. These two items make keyed access impractical since the extreme length of the key would negate any benefit derived. The only alternative to keyed access is sequential access, but this is undesirable because of the time required to search through the large volumes of information generally contained in a matching database.

So, we are presented with no other alternative but to somehow reduce the size of the key, thus making keyed access viable. There are many well known data compression techniques which could be used to do this, a general survey of which can be found in Reghbati (1981). In ACTR, the required data compression is achieved by forming the "compressed phrase key" or CPK. How CPK's are actually formed is discussed below. Accept for now that CPK formation results in a key which is approximately 35% of the original size of the phrase. The CPK can thus be used to access the matching database with an efficiently sized key in order to determine whether any direct matches exist.

The use of the CPK in ACTR is significant in the following ways:

1. All 100% matches will always be located using this method.
2. Since ACTR is able to locate direct matches by using the most efficient means possible, matches made by using this method are both faster and cheaper to perform.
3. As applications mature, the proportion of direct matches generally increases due to ongoing database update activity on the part of the user. Thus, overall matching costs for an application can actually decrease as the application matures, even though the size of the matching database may increase.

CPK Formation

The CPK is formed by first ordering the words defined in parsing. The actual order is arbitrarily chosen and so is not significant, as long as the same ordering applies for all CPK formations. (The order used happens to be in ascending order of the collating sequence in use.)

After ordering, the words are concatenated into a single string which contains no blanks. This string is then compressed in order to form a short enough string to allow for efficient use as a database retrieval key. The compression of the string is based on the following:

1. The words resulting from parsing generally contain only characters from the 26 alphabetic character set and the 10 character numeric set. (Recall that the actual set of characters which may be encountered in words is user-defined.) However, characters are stored internally (*ie.* in memory and on disk) using an 8 bit code. Thus, there are 2^8 or 256 possible 8 bit code combinations while ACTR words typically use no more than 36 of these. This leaves a 220 code surplus which could be used for other purposes.
2. Certain double and triple letter combinations are known to occur more frequently than others in English and French text samples. In ACTR, the double letter combinations are known as "digrams", and the triple letter combinations are known as "trigrams".
3. The 220 "free" codes can then be used to replace the digrams and trigrams described above as they occur in text samples.
4. Starting with the concatenated, parsed words, ACTR scans for the presence of any of the predefined digrams and trigrams. If any are found, they are replaced with the associated 8 bit code. The result is that a character sequence which formerly required 16 or 24 bits of storage, now requires only 8 bits.

Indirect Matching

Like direct matching, indirect matching begins with the set of words resulting from the parsing process. However, indirect matching can never be as efficient as direct matching since the concept of closest match is relative. That is, we cannot find the closest match without first performing an exhaustive search through all of the possible matches.

In order to perform indirect matching, the matching database must first be searched for each of the words resulting from the parsing process in order to determine which, if any, are known. Following this step, for each word in the supplied phrase which is known to the database, all phrases containing the word must be retrieved and evaluated.

The nearest matching phrase is determined by calculating a score for each of the possible matches. Scores are based on the weights of the words which are in common with the database and subject phrases. Of all database phrases evaluated in this manner, the highest scoring phrase is the one which is considered to be the closest match.

Word Weight Calculation

For each word known to the database, ACTR calculates a matching heuristic, or weight. These weights are an indication of the usefulness of a word in assigning a code and act as components in the phrase score calculation process.

The method by which word weights are calculated is based on: n , a count of unique codes, whose associated phrases contain this word; V_i , the relative frequency of code i from previous surveys; X_i , a count of the number of word occurrences for phrases with code i ; P_i , the proportion of this word in code i , calculated as $V_i \times X_i / \sum_{j=1}^n V_j \times X_j$; EW , the entropy of the word, calculated as $-\sum_{i=1}^n P_i \times \log_2 P_i$; K , the total number of word occurrences for code i , calculated as $\sum_{i=1}^n X_i$; EU , the entropy of a uniformly distributed variable with K unique values, calculated as $\log_2(K)$; and finally EO , a small value to avoid division by zero, calculated as $-K/K + 1 \times \log_2 K/K + 1$.

From the preceding, word weights are calculated as: $EU - EW + EO/EO + EW$.

Phrase Score Calculations

For each database phrase which is evaluated for an indirect match, a score is calculated. The score is based on: n , the number of words the phrases have in common; w_k , the weight for word k ; m , the number of words in the subject phrase; and l , the number of words in the database phrase.

From the preceding, phrase scores are calculated as: $n^3 \times \sum_{k=1}^n w_k / m \times l$.

Matching Parameters

After calculating a score value for each potential match, ACTR compares the score against user supplied values for the following parameters and takes the action indicated.

1. UPPER THRESHOLD

If the resulting score is greater than or equal to this value, then a winner is considered to have been found.

2. LOWER THRESHOLD

If the resulting score is greater than or equal to this value, but less than that supplied for the upper threshold value, then a possible match is considered to have been found.

3. PER CENT DIFFERENCE

If more than one winner is found, and their scores are within the supplied value for this parameter, then multiple winners are considered to have been found.

Limiting the Search for an Indirect Match

ACTR searches the matching database for possible matches using the known words in the subject phrase. That is, these words are used to search for database phrases which contain them. The search proceeds in order of the ascending frequency of occurrence of the known words. Thus, the known word which occurs the least frequently in the database is used to start the search, the next lowest is used to continue the search, and so on.

As can readily be appreciated, finding a match by the indirect process has the potential of being time consuming and very expensive. Unfortunately, attempts to find matches by indirect means are unavoidable since a nearest matching feature is an essential component of any automated coding system.

While performing a search in this manner, ACTR maintains a list of database phrases which have already been evaluated. After a database phrase has been evaluated, it will not be re-evaluated in a subsequent iteration for the currently executing matching effort. This ensures that a database phrase which contains more than one of the known words will not be evaluated more than once.

As a further search optimization, ACTR makes use of the user supplied matching parameters. With these, it constructs a table of optimistic scores for each iteration of the word based search:

1. For the first known word, the optimistic score is based on the possible occurrence of a database phrase with the same number of words as the number of known words and with all of its words in common with the subject phrase's known words.
2. For the second word, a similar assumption is made, but since the first word has already been used in the preceding search iteration, we know that any phrase containing the first word has already been evaluated. So, the optimistic score is based on the presence of the second and subsequent words only.
3. Optimistic scores for succeeding iterations are based on the presence of the current and succeeding unsearched words only.

The formula used to calculate the optimistic scores is based on: a , the number of known words in the subject phrase; b , the number of words in the subject phrase already searched; c , the total number of words in the subject phrase; and d , the number of known words not yet searched, calculated as $a - b$;

From the preceding, optimistic phrase scores are calculated as: $(d^2 \times \sum_{i=d}^a w_i) / c$.

With the table of optimistic scores in place, ACTR evaluates the potential score at each iteration before performing a database access. Thus, hopeless searches are never attempted.

To summarize, the search for an indirect match is terminated when any of the following conditions are met:

1. The maximum potential score for the current iteration does not meet or exceed the threshold defined for possible matches.
2. At least one match has been found and the maximum potential score for the current iteration cannot produce another.
3. The maximum number of possible matches requested by the user has already been found and the maximum possible score for the current iteration does not exceed that of the lowest scoring phrase.

5. SUMMARY

A flexible and efficient automated coding methodology, embedded in a generalized software system has been presented. The system can be used to perform automated coding for any application in English or French or both, using any classification scheme. In doing so, it makes use of a powerful generalized parsing strategy and significant performance optimizations. For further information on ACTR, the interested reader is directed to Connor, Salloum and Wenzowski (1988).

6. ACKNOWLEDGEMENTS

While a complete list of all those who have been involved in the ACTR project would be too large to be presented here, the author would specifically like to acknowledge the contributions made by: John Connor and Bill Salloum, the principal programmers for the project; Victor Estevao, who was instrumental in helping to design the parsing strategy; Malvinder Rakhra and Paul Surman, who performed a great deal of system testing; and Don Royce, who was the project manager for the research effort.

REFERENCES

- APPEL, M., and HELLERMAN, E. (1983). Census bureau experiments with automated industry and occupation coding. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 32-40.
- CONNOR, J., SALLOUM, B., and WENZOWSKI, M. (1988). ACTR Documentation Set. (System Overview, User's Guide, Tutorial, Guide to the Parsing Strategy, Default Parsing Data, Message Guide, Command Language Guide, Searching and Matching Methods & Programmer's Guide) Internal Documents, Statistics Canada, Research and General Systems Subdivision, Ottawa, Canada.
- LANDRY, L., and PIDCOCK, J. (1984). Business Register Automated SIC Coding System, System Proposal and Design. Internal Document, Statistics Canada, Informatics Services and Development Division, Ottawa, Canada.
- REGHBATI, H. (1981). An overview of data compression techniques. *Computer*, 14,4, 71-75.

Quality Control Processing System for Survey Operations¹

WALTER MUDRYK²

ABSTRACT

The methods used to control the quality of Statistics Canada's survey processing operations generally involve acceptance sampling by attributes with rectifying inspection, contained within the broader framework of Acceptance Control. Although these methods are recognized as good corrective procedures, they do little in themselves to prevent errors from recurring. As this is of the utmost importance in any quality program, the Quality Control Processing System (QCPS) has been designed with error prevention as one of its primary focuses. Accordingly, the system produces feedback reports and graphs for operators, supervisors and managers involved in the various operations. The system also produces information concerning changes in the inspection environments which enable methodologists to adjust inspection plans/procedures in accordance with the strategy of Acceptance Control. This paper highlights the main tabulation and estimation features of the QCPS and the manner in which it serves to support the principal quality control programs at Statistics Canada. Major capabilities from a methodological and systems perspective are discussed.

KEY WORDS: Quality control processing system; Process control; Acceptance sampling; Acceptance control; Skip-lot sampling.

1. INTRODUCTION

This paper deals primarily with the features of the Quality Control Processing System (QCPS) that is presently being used at Statistics Canada. However, in order to show how this system fits into the overall quality picture for surveys, the paper begins with a brief discussion of the survey process and the role that quality assurance and quality control play in this process. The paper then identifies the specific quality control methods and strategies that are used for processing operations at Statistics Canada and how the QCPS serves to support this activity. The paper then proceeds to describe the system features and provides a summary of its major achievements.

1.1 The Survey Process

The requirement of ensuring quality in the overall survey process has always been considered a high priority at Statistics Canada. In a very general sense, it may be viewed as being achieved through the application of a series of quality assurance (QA) and quality control (QC) measures at the appropriate stages of a survey process. It is important to distinguish between these two activities since in our environment, they involve very different approaches and procedures that are normally applied at different points in the process. A simplified overview of the survey process at Statistics Canada includes the following stages:

¹ This is a revised version of the paper presented at the Fourth Annual Research Conference, Bureau of the Census, Arlington, Virginia, USA, March 1988.

² W.V. Mudryk, Business Survey Methods Division, Informatics and Methodology Branch, Statistics Canada, 10-J, Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6.

- planning
- design
- implementation
- processing
- publication.

It is important to note that every one of these stages is subject to some error. It should also be realized that the further into the survey process the errors are discovered, the more impact they have on survey timeliness, cost and accuracy. Therefore, it is good practice to put a strong emphasis early in the process, on the development of measures and procedures that would prevent or reduce their occurrence. This should occur at the planning and design stages of the survey process. These measures and procedures are also known as quality assurance.

1.2 Quality Assurance

A general approach to establishing quality assurance is to try to anticipate problems very early in the survey process and take appropriate steps to prevent or minimize them. The anticipation can be based on experience, reviews, evaluations, debriefing exercises, feasibility studies, *etc.* The steps could include improving sampling frames/designs, modifying data collection methods, improving questionnaire design, providing clearer processing procedures, *etc.* A comprehensive list of such steps may be found in Statistics Canada's Quality Guidelines (1987).

This approach is extremely important since effectively it moves quality upstream and thereby helps to prevent many potential problems from occurring. Furthermore, in so doing, it assures better quality at the least cost by "getting it right the first time". Despite our best efforts however, there are some situations when error levels continue to be unacceptably high. In these situations we consider the use of quality control.

1.3 Quality Control

In contrast with QA, statistical quality control has been found to be highly applicable at the processing stage of the survey cycle. At this stage, the work usually has the following characteristics:

- labour intensive and repetitive in nature;
- assigned to individuals or operators with varying abilities;
- normally grouped into batches or lots of similar work units.

As such, these survey operations are more prone to the occurrence of errors. Examples of these operations include:

- coding/transcription
- manual editing/reviewing
- data capture/entry
- corrections/reconciliation
- updating/profiling, *etc.*

For many reasons, which include complexity of tasks, abilities of operators, turnover of staff, *etc.*, the amount and significance of error varies between operations, between operators within an operation, and at times within operator. Statistical quality control is used to identify and reduce this variability and ensure that the outgoing quality of each operation falls within acceptable levels.

2. QUALITY CONTROL STRATEGY

2.1 Methods of Quality Control

Of the two main methods of quality control available, namely, process control charts and acceptance sampling, we have found the latter methodology applied in the broader context of Acceptance Control, to be the more appropriate method for *on-line* quality control of survey processing operations. The reasons for this are as follows:

- prior control or stability of process cannot be assumed initially nor always attained in the long run;
- assignable causes of error are not always known since we are dealing with people (vs. say machines);
- processes cannot readily be stopped and adjusted for assignable causes, even if they are known;
- with many operators and large “between operator” variabilities, many individual control charts requiring immediate updating (*i.e.*, after each sample observation) would be required on-line to the survey operation; this would be operationally difficult to achieve.

Therefore our quality control strategy generally consists of using varying acceptance sampling procedures (with rectification) applied at the operator level, as a screening device for correcting substandard quality, with the aim of continually reducing inspection as the inspection results support this action. This is coupled with an emphasis on operator and supervisor feedback to establish error prevention. In this manner both error correction and subsequent prevention are exercised at the error source, where they can have their greatest impact. Furthermore, between operator variations are automatically dealt with as each operator is effectively treated as a process in the following sense. During a period of low to moderate stability, acceptance sampling is applied to each lot processed. During a period of high stability coupled with good past inspection results, less acceptance sampling and even spot checking may be applied under the broader strategy of Acceptance Control.

2.2 Acceptance Control

After a quality control program has been operating for some time, operator processing abilities tend to improve and in many cases, a stabilization of quality occurs. In an effort to take advantage of this improved situation and to enable our quality control designs to be more economical, we have adopted the strategy that Schilling calls Acceptance Control (1982). Under this approach, acceptance sampling procedures are continually modified and adapted as changes in the inspection environment are identified. This is in accordance with one of QC's main pioneers, H.F. Dodge who states (1950):

“A good product with a history of consistently good quality requires less inspection than one with no history or a history of erratic quality. Accordingly, it is good practice to include in inspection procedures provisions for reducing or increasing the amount of inspection, depending on the character and quantity of evidence at hand regarding the level of quality and the degree of control shown.”

In fact the ultimate aim of acceptance control is to continually reduce inspection to the level of spot checks or process controls as the quality history improves and stabilizes. At Statistics Canada, two specific approaches are used to achieve this principle:

- *Graduated Inspection Plans*. These are obtained by raising or lowering the quality index for the sampling plan as changes in the process average are observed and then closely monitoring the impact on the resulting average outgoing quality estimates.
- *Cumulative Results Plans*, more specifically Skip-Lot Sampling (Stephens 1982). Here, the extent of skipping lots depends on the stability and level of expected incoming quality.

Both approaches are part of our acceptance control strategy and require a good quality history which would indicate not only the underlying level of processing quality (*i.e.*, at the operator level) but also the extent of stability (*i.e.*, degree of control) that can be expected in the process. Accordingly, the inspection process must provide:

- good data (accurate error estimates);
- quick results (monthly, weekly, daily);
- incentive for improvement (feedback reports);
- quality history (time series of error quality).

Essentially these have been the motivating influences in developing the Quality Control Processing System (QCPS). It should be noted that changes are currently being made to the system to expand the existing operator quality history. This should provide the data to enable greater implementation of spot checks and/or process control for selected operators with exceptional and stable performances.

3. SYSTEM DESCRIPTION

Based on the strategy identified above, the QCPS has been developed to achieve the following objectives:

- process any single acceptance sampling transaction;
- provide output by operator where each operator can be treated as the error source;
- provide feedback to four levels of staff with current and historical quality control information;
- support the acceptance control strategy by enabling the processing of skip-lot sampling results and providing an extensive operator quality history;
- support the major QC objectives of error correction and prevention while enabling inspection costs to continually be minimized.

3.1 Methodological Features

a. Inspection Schemes

The system can process any quality control transaction resulting from the application of *single* acceptance sampling. This naturally includes normal, reduced and tightened plans as well as any skipped lots resulting from skip-lot sampling. The system will also process any lot whose plan designation is 100% inspection.

b. Lot Status Codes

The system determines the treatment of incoming QC transactions by using lot status codes which indicate the state of completeness of the intended inspection. There are codes for the following lot situations:

- sample inspected and accepted;
- sample inspected and rejected (remainder inspected);
- 100% inspected;
- any of the above not completed (3 codes);
- no sample inspection due to skip-lot.

c. Attributive Quality Measures

The system will produce estimates for various quality measures which include percent defective, defects per hundred units and weighted error equivalents. For the latter quality measure, the system allows errors to be weighted according to a pre-defined error seriousness classification scheme. Typically, under these more complex measures, errors are categorized and assigned weights from 0 to 1 depending on their relative magnitude and seriousness. For purposes of simplicity, no more than four error categories are generally defined, as follows:

Category	Weight
Critical	1.0
Major	0.4 – 0.6
Minor	0.2 – 0.3
Insignificant	0.0 – 0.1

d. Estimates

The system provides estimates and their associated standard errors (where applicable) for many key quality control indicators. The most important of these are:

(i) Error Rates

Error rates are calculated which relate to the individual operator, a specific sampling plan or the overall application. These estimates are provided for various time frames (*e.g.*, daily, weekly, monthly, quarterly, *etc.*), and various subsets of the application, such as specific lot categories (*e.g.*, rejected lots) or sub-groups (*e.g.*, regional offices).

(ii) Operator Process Average

An estimate of an operator's processing ability at any particular point in time is provided by the operator process average. This estimate is calculated using an empirical Bayes approach (MacMillan and Mudryk 1988) which essentially shrinks the current operator sample error rate estimate part way towards the grand average error rate of the last four periods for that operator. The basis of shrinkage is determined by the ratio of the sampling variance of the current sample estimate to the total variance of the grand average estimate. This quantity has been found to produce good estimates for qualifying operators onto minimum inspection sampling plans.

(iii) Rejection Rates

Actual and expected rates of rejection are calculated for each sampling plan for purposes of statistical comparison and operational evaluation. The expected rates are obtained assuming Poisson probabilities.

(iv) Inspection Rates

Inspection rates are calculated at various levels as a general indicator of relative costs. These rates are determined with and without skip-lot effects on an actual and expected basis. The expected rates are a natural extension of the expected rejection rates discussed above.

(v) Average Outgoing Quality

An estimate is provided of the Average Outgoing Quality (*i.e.*, AOQ) rate resulting from the application of quality control to the operation. This estimate projects the observed error rate at the operator level to the uninspected volume for that operator, and then aggregates all operators to determine the overall estimate.

e. Analysis

The system provides tabulations and outputs which enable analyses to be performed at various levels which help to subsequently fine tune the application parameters and/or modify the plans. These include:

- operator profiles that enable a sampling plan/procedure qualification analysis;
- individual sampling plan evaluations that provide an overall QC plan analysis;

- summaries of key indicators that enable a QC cost-benefit analysis;
- a Pareto analysis of operator and error code contributions;
- group charts of operator process averages that provide an operations performance analysis.

f. Reports

The system produces 8 reports and 5 graphical outputs (through its link to SASGRAPH) for each application run. Tabulations can also be produced for specified sub-groups (*e.g.*, Statistics Canada's regional offices) with a summarizing feature over all sub-groups of each report.

Each set of output reports is designed for and disseminated to four levels of staff, namely: operator, supervisor, manager and QC designer. Examples of the output reports are available from the author.

3.2 Software Features

a. Operator Capacity

For each application, the system can handle up to 108 operators in its historical file, each containing up to three previous periods of error information. A unique self-maintaining feature of this file is that any operator who has not been active during at least one of the last 4 consecutive months of processing is dropped. This makes room for new operators on the file and thereby increases the effective file capacity.

b. Historical Updates

The system updates each operator error quality history (of up to 4 consecutive periods) with new information as it becomes available. This is currently being increased to 6 consecutive time periods. If an operator has not processed during a particular month, blank data for that month is inserted. Likewise, application year-to-date and quarterly totals are updated with the addition of each new month of QC data.

c. Year-End Rollover

Most of the QCPS applications are maintained on a calendar year basis. When this option is specified, the system will zero out the previous monthly totals and commence a new application time series (usually starting in January). The quarterly totals and the operator error quality time series however, are not re-set at this time and continue to be maintained as usual.

d. Recovery

If a tabulation run is made and errors are subsequently discovered, another run can be made using the recovery feature with the corrected data, to automatically produce the corrected outputs.

4. SYSTEM BENEFITS

The QCPS is aimed at servicing the needs of four levels of staff which interface with each QC application. Accordingly, the major achievements of this system can best be described under these same headings:

a. Operator Level

The QCPS provides extensive feedback to the individual processing operators on their current and historical performance. The operators are then able to track their own progress, compare their own performance with that of their peers, and identify explicitly where their errors are being made. The result of this feedback generally leads to:

- improvement in operator processing ability;
- increased motivation with respect to peers;
- greater quality consciousness;
- higher operator morale.

b. Supervisor Level

The system provides operational information to the supervisors which enables them to better manage their operation in terms of:

- effective resource allocation and work distribution;
- identifying problem operators and/or areas;
- determining training needs.

c. Management Level

The system provides data summaries on key quality control indicators for management which enables them to:

- receive an assurance of quality;
- track the progress of the application in terms of quality and costs;
- recommend changes to operational objectives.

d. QC Design Level

The system provides extensive information (*e.g.*, estimates, quality histories) which is used to analyze the quality control design and fine tune or enhance the methods and procedures of each application. When this data has been established and maintained over a sustained period of time, it can lead to:

- improvements in QC methodologies and procedures;
- sampling plan and/or inspection procedure adjustments;
- minimization of inspection costs.

5. CONCLUSIONS

The QCPS is being used at Statistics Canada to support the Quality Control programs of many production oriented survey processing operations. As the ultimate aim of each program is to exercise error prevention to the extent possible, as well as, to progressively reduce inspection to the level of spot checks, a good and flexible processing system is essential. The QCPS achieves these objectives by providing good data and quick results to the various levels of staff that are involved in each operation, as well as, supporting the various inspection methods that fall under the general strategy of Acceptance Control.

The system is particularly attractive to our user community since it can easily handle large volume operations involving many operators, quickly and at a low cost. Furthermore, by treating each operator individually, the system focuses attention to each relevant error source and supports this with necessary feedback to the appropriate levels of staff. In this manner the system enables our quality control methods to be both preventive and corrective in an efficient and economical manner.

ACKNOWLEDGEMENTS

The author would like to thank the referees and Jeffrey Smith for their constructive comments in reviewing this paper. Examples of output reports relating to this system are available from the author.

REFERENCES

- DODGE, H.F. (1950). Inspection for quality assurance. *Industrial Quality Control*, 7(1), 8.
- MacMILLAN, J.H., and MUDRYK, W.V. (1988). A non-parametric Empirical Bayes approach for estimating a process average in quality control. Paper presented for the Section on Physical and Engineering Sciences, American Statistical Association Annual Meeting, New Orleans, Louisiana.
- MUDRYK, W.V., and BOUGIE, R.W. (1987). Quality control processing system (QCPS) - Users Manual. Internal document. Statistics Canada, Ottawa.
- SCHILLING, E.G. (1982). *Acceptance Sampling in Quality Control*. New York: Marcel Dekker.
- STATISTICS CANADA (1987). *Quality Guidelines*. 2nd Edition, Ottawa, Canada.
- STEPHENS, K.S. (1982). *How to Perform Skip-Lot and Chain Sampling*. Volume 4, ASQC Basic References in Quality Control: Statistical Techniques. American Society for Quality Control, Milwaukee, Wisconsin.

Postal Address Analysis

YVES DeGUIRE¹

ABSTRACT

When we examine postal addresses as they might appear in an administrative file, we discover a complex syntax, a lack of standards, various ambiguities and many errors. Therefore, postal addresses represent a real challenge to any computer system using them. PAAS (Postal Address Analysis System) is currently under development at Statistics Canada and aims to replace an aging routine used throughout the Bureau to decode postal addresses. PAAS will provide a means by which computer applications will obtain the address components, the standardized version of these components and the corresponding Address Search Key (ASK).

KEY WORDS: Postal addresses; Administrative data; Parsing; Standardization; Search key.

1. INTRODUCTION

Postal address analysis can be defined as the process of identifying the basic components of an address which appears in free format, standardizing those components, and generating an identifier for that address. This process can be used, for example, in the pre-processing step of any record linkage application that uses an address field or in the generation of a key for database access. Statistics Canada, as part of its 1991 census research program, is conducting a study on the implementation of a national Address Register. Such a register contains basically, postal address information. This information must be analyzed carefully in order to produce a register and to assess its quality. The Address Register Research Team has recognized that fact and research into the area of automated postal address analysis was initiated.

This paper presents the results of this research on postal address analysis. The nature of an address and its related problems will be described. Also, some computer considerations will be discussed to explain why new software is needed for the Address Register and Statistics Canada. Finally, we will examine PAAS (Postal Address Analysis System); a system currently under development at Statistics Canada.

2. POSTAL ADDRESSES: STATEMENT OF THE PROBLEM

A postal address can be defined as a string of characters representing a location where an individual can pick up his mail. By location, we mean a physical place where the deliverer (like a postman) and the receiver agree in the matter of mail reception. It can be a dwelling, a postal box, a street or a rural route. To restrict our field of study, we are going to examine the addresses that are Canadian (French and English), that represent residential locations and that should result in correct mail delivery.

¹ Yves DeGuire, Research and General Systems, Statistics Canada, room 2405, Main Building, Tunney's Pasture, Ottawa, Ontario, K1A 0T6.

As one would expect, the flexibility in the address definition results in problems for any computerized application having to deal with postal addresses. Even a person is likely to encounter some problems with addresses with which he/she is not familiar. Three major problems are analyzed here.

2.1 The Syntax of a Canadian Postal Address is Complex

A postal address is composed of tokens (lexical items which can be considered as basic units in an address). A token can be either a delimiter, a term (or keyword), a word, a letter or a number. Figure 1 illustrates an example of token decomposition. Tokens can be combined to get address components which are larger address structures. In turn, a component can fall into three groups: designators, qualifiers and secondary words. Figure 1 gives also an example of a component decomposition. Valid addresses are composed of both a set of valid combinations of components and a set of valid combinations of tokens. However, it is more practical for implementation purposes, to define an address with token patterns (combinations of tokens). Token patterns can be generated from a formal postal address grammar (written in BNF for example) and used directly for constructing a postal address.

This syntax is fairly complex. First of all, the grammar is sizeable. We have analyzed a national sample of 30,000 addresses taken from six different administrative files. In these addresses, we found around 4,900 different token patterns. This is substantially higher than what is reported in Drew(1987) because we have analyzed addresses from many different files, not just one. Other interesting results concern the distribution of those patterns. Only 37 patterns are necessary to cover 50% of the addresses. So, there are a few common patterns, but most of the patterns are rather rare. Nevertheless, this analysis illustrates the complexity of postal address syntax by demonstrating that it is not restricted to just a few patterns. Secondly, as much as 600 different terms can be found in a good national sample of addresses. Thirdly, an address is usually in free format, *i.e.* the components (and the delimiters) can occur in any one of several positions.

2.2 Addresses Don't Follow Precise Standards

Addresses representing the same address location can be written in many ways as illustrated in Figure 2. The reason for this situation is the flexibility in postal address syntax and also human nature. In fact, people write addresses as they like and follow the "standards" in use in their immediate environment.

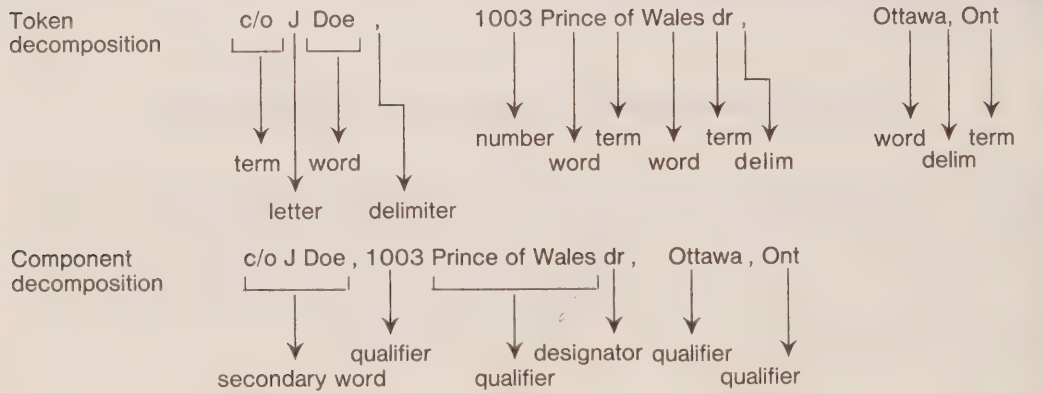


Figure 1. Two ways of decomposing a postal address.

2.3 Ambiguities Occur in Postal Addresses

A postal address can't be regarded only from a syntactic point of view. Its semantic (i.e. the meaning a postal address) must be examined as well. Sometimes, one address can potentially represent more than one location. We then face an ambiguity since we don't know how to interpret it. To do so, more knowledge is required in order to exclude the locations that don't exist and to identify the correct location. However, this knowledge doesn't always permit us to narrow down the location; we then face an unresolvable ambiguity. Figure 3 shows an example of an ambiguous address.

3. COMPUTER SYSTEMS CONSIDERATIONS

Now that we have a better understanding of postal addresses as well as their related problems, we will concentrate on the use of postal addresses in computer systems.

3.1 Computer Applications Requiring Address Information

Several types of application require address information. Some record linkage projects link individuals or dwellings (like in the construction of an Address Register) based on their postal addresses. Their linkage rules perform essentially on standardized address components. On the other hand, databases and computer files storing postal addresses are numerous. For example, postal addresses information for an Address Register must be stored in some fashion, either in a stand alone flat file or in some kind of integrated database. But what information is stored? Address components (standardized or not) could be. For follow-up or historical purposes, the original input address could be kept as well. However, retrieval from a large database (or a large flat file) requires an Address Search Key (ASK) to allow direct access (or direct matching) to a record identified by a postal address. Mailing labels processing is another area where postal addresses is a big concern. Address components, standardized or not, can form mailing labels.

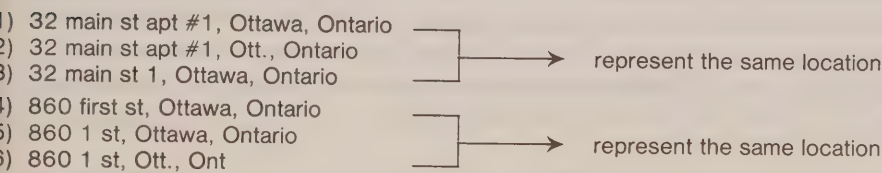


Figure 2. Examples of Addresses Which Represent the Same Location.

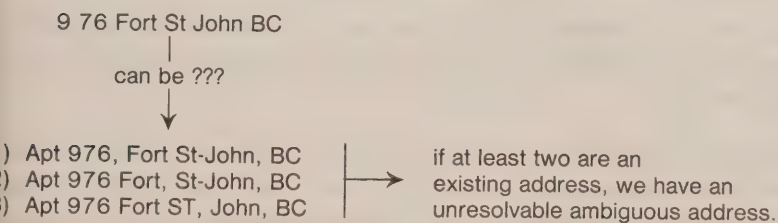


Figure 3. Example of an Ambiguity.

3.2 Three Basic Information Components

Therefore, three basic information components need to be derived from a free format postal address: the address components, the standardized components and the Address Search Key (ASK).

1. THE ADDRESS COMPONENTS

They represent recognizable and useful portions of an address. The major address components are street number, street name, street direction, street designator, postal designator, postal qualifier, municipality name, province name, and postal code.

2. THE STANDARDIZED COMPONENTS

They are the standardized version of the address components, where any style variations are removed.

3. THE ADDRESS SEARCH KEY (ASK)

This is a compressed string, unique for a given address.

3.3 Postal Address Analysis System

A complete Postal Address Analysis System (a computer system that generates the three basic information components we need) represents an expert system in the field of postal addresses. Expert because you replace a specialist (like a postman) in address recognition. At Statistics Canada in the 1970's, two routines were developed to analyze postal addresses. ENCODA (component decomposition) and ASKGEN2 (standardization and ASK) were implemented for the Business Register Maintenance System. They served well until recently. With the advent of powerful computers, new software development techniques and the Address Register itself, they don't perform to today's standards.

- The encoding success rate is too low. A study using a national sample of addresses from many administrative files shows that ENCODA cannot properly decode an address, on average, 15% of the time. This is not acceptable since it could lead, in the case of the creation of a national Address Register, to over one million encoding failures.
- The user interface is poor. There is no comprehensive status produced at the completion of the analysis. As well, very few utilities are provided in order to ease programming burden.
- The functionality is incomplete. Standardized components and ASK are mixed up in the same data structure. Standardized components are truncated to allow data compression but ASK is very long because it is stored in fields of fixed length. Also, the software doesn't recognize address ambiguity.
- Maintenance of the software is a nightmare. New address patterns are difficult to incorporate into the routines because these are complex and tend to become more and more so with time. This is a sign of aging software.

To fulfill the requirements of an Address Register and of Statistics Canada in the area of address analysis, the development of a completely new system was initiated. The problem this time has been approached with expert system techniques, modular design and full scale implementation. This new system is called PAAS, for Postal Address Analysis System.

4. A POSTAL ADDRESS ANALYSIS SYSTEM: PAAS

PAAS is currently under development. Therefore, some results are preliminary, but in general very encouraging. We will review here the four basic functions of the system.

4.1 Address Parsing

The parsing function is the most important and complex function of PAAS. Here, PAAS accepts as input a free format address, scans it (breaks it into lexical items) and parses it (analyzes the syntax) to decode it into address components.

This parser generates the following items for every address processed (Figure 4 illustrates two examples of this output):

- A comprehensive Address Status code; such as V for valid, E for syntax error, *etc.*
- Identification of components in the input address.
- Components classification: every component is classified using a detailed code, so it is easy to understand the meaning of a component. This code is divided into three sub-codes:
 - TYPE code: indicates the group of components to which a component belongs. Example of TYPES are those for province (PR), municipality (MU), street (ST), *etc.*
 - CAT code: refines the group of components indicated by TYPE. Examples for the street TYPE (ST) are name (NA), number (NU), designator (DE), *etc.*
 - CLASS code: classifies a component by examining its characteristics. Examples are avenue (AV) or road (RD) classification of a street designator.
- Ambiguity detection: the PAAS parser flags any component that could change because of an ambiguity.

The PAAS parser was implemented using MPL. MPL is a meta-programming language. It allows us to generate programs or subroutines used for syntax analysis and automatic translation. The input to MPL is a set of specifications divided into the scanning (token recognition), the syntax rules and the semantics. The scanning represents the lexical analysis where the input is broken down into tokens. The syntax specification is similar to a BNF grammar specifications: the right-hand side symbols of a syntax rule are defined by the left-hand side symbols. Figure 5 gives examples of syntax rules. Finally, a semantic action can be associated with any rule and is used to handle some complex aspects of the syntax, as well as to perform other actions (such as updating a table of components). The MPL language is well suited to writing translation specifications and has been used at Statistics Canada to implement STATPAK (retrieval

ADDRESS	COMPONENT	TYPE	CAT	CLASS	AMB_FLAG
(1) 32 Main st, Ottawa, Ont	32	ST	NU	**	
	Main	ST	NA	**	
ADDRESS__STATUS====> V	st	ST	DE	ST	
	Ottawa	MU	NA	**	
	Ont	PR	NA	35	
(2) 32 Main st Ottawa Ont	32	ST	NU	**	
	Main	ST	NA	**	
ADDRESS__STATUS====> A	st	ST	DE	ST	*
	Ottawa	MU	NA	**	*
	Ont	PR	NA	35	

Because the second example misses the commas to delimit the address, an ambiguity is flagged by PAAS.

Figure 4. Examples of the PAAS Parser Outputs.

and tabulation system for the census), NYSIIS (name encoding routine) and NAMEPARS (name parser). It saves development time (*e.g.* you don't need to write a detailed and custom program in a traditional programming language such as PL/1). The specifications in BNF are much easier to understand than is a program with a complex logic.

The PAAS parser involves a rather complicated syntax analysis and represent a fairly important MPL application. For example, a dictionary containing more than 600 terms assist in the scanning of addresses. As well, more than a hundred syntax rules implement the syntax analysis. In this syntax analysis, the initial tokens are transformed from a rule right-hand side to a rule left-hand side and become higher level address fragments (this is known as forward chaining) until the address is completely analyzed. During this process, the address components are identified and stored in a table by the semantic action of a rule. The invalid addresses are found whenever no rule is applicable. A sample set of rules to decode an address is illustrated in Figure 5. Finally, for some complex addresses, a special analysis is performed through the use of the MPL semantic facility. This is required anytime an ambiguous term is encountered. In this case, PAAS analyses the surroundings of the ambiguous term.

In comparison with ENCODA, the PAAS parser is an improvement in the following are as:

- The quality of the parsing: the PAAS parser is able to decode more addresses successfully than ENCODA does. A series of parallel runs over identical national samples of addresses showed that PAAS is successful on more than 97% of addresses, while ENCODA properly handles only 85% of them.
- The indication of an address status: the status is more complete than ENCODA's which provides for only two possibilities: decoded address or blank address!
- The components: PAAS generates much more comprehensive component information than does ENCODA.
- The maintenance: the utilization of MPL helps in making the PAAS parser a lot easier to maintain than a huge algorithm such as is used by ENCODA.

4.2 Components Standardization

The standardization aims to remove any style variation in the address components defined in the parsing phase.

Unlike ASKGEN2, PAAS doesn't truncate any component and retains all the information in the components. This standardization is achieved basically in three different ways depending on the nature of the component:

1. CODABLE COMPONENTS

Every component for which a limited number of values exist is standardized by replacing its value with the CLASS code of the component (this code uniquely identifies the standardized value of the component). Falling into this category are components such as the province name, street designator, *etc.*

2. NAME COMPONENT NOT NUMBERED

To standardize a non-numbered name component, several rules must be applied to transform the original value into a standardized value. The rules vary from the removal of useless characters (*e.g.* quote, hyphen, *etc.*) to abbreviation replacement (*e.g.* Mtl becomes Montreal).

3. NAME COMPONENT NUMBERED

A numbered name component is standardized by returning its name as a number. For example First becomes 1, Second 2, *etc.*

The Address Search Key should be unique and short.

Address to parse: 100 Rideau st Ottawa Ont K1N5X2

At some point, we have a string of address fragments which will be transformed by five rules. The “|” denotes a “OR” and [] is an optional syntax element.

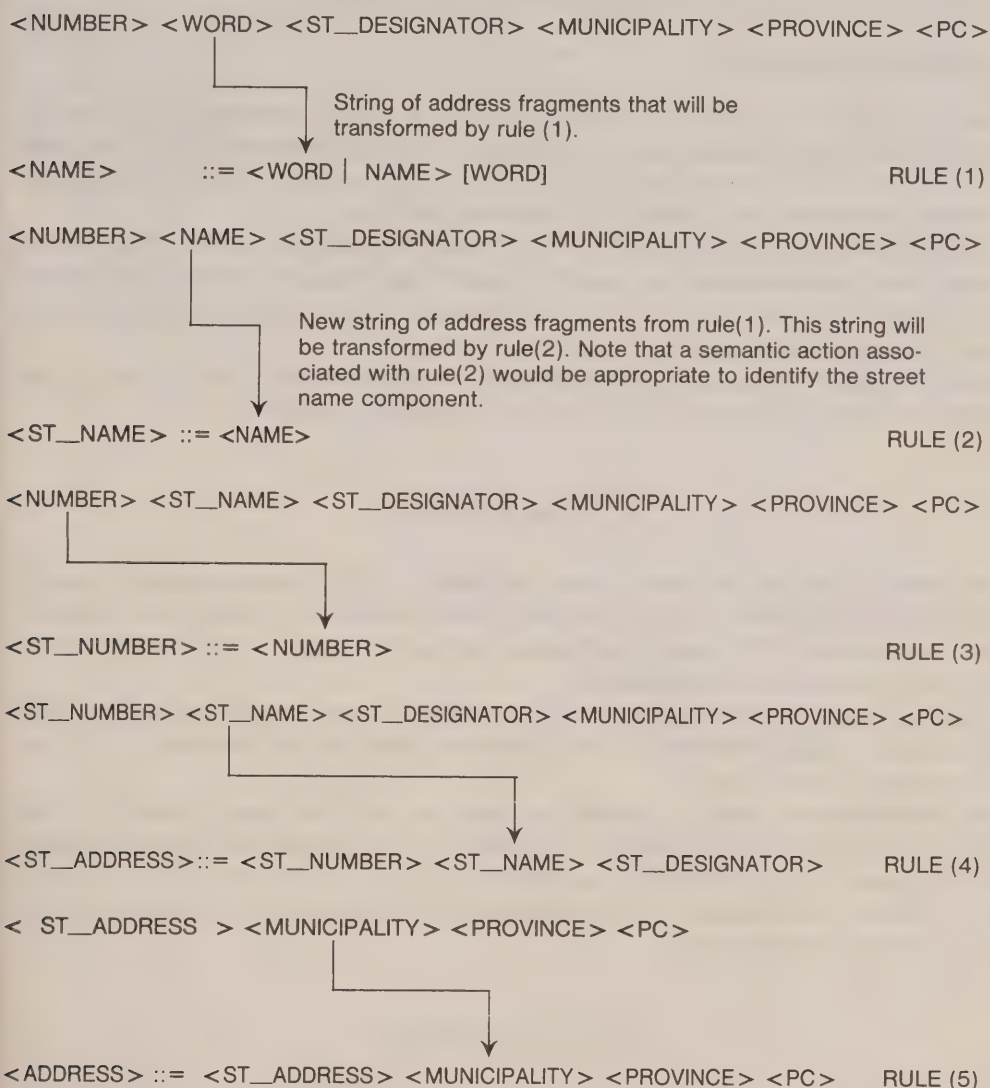


Figure 5. Rules for a Sample Address Syntax.

To shorten the key, different compression techniques can be used. However, compression takes time and we have to choose a technique that will be efficient. We are experimenting with two different techniques.

1. TRUNCATION

here, the name components are truncated. This technique is not real compression and could affect the uniqueness of a key. However, it is simple and fast.

2. REAL COMPRESSION

a compression technique that we are looking at consists basically of replacing common combinations of characters by a character code not in use for writing an address. Here, we will preserve the uniqueness but increase the complexity of generating and using a key. Therefore, a longer ASK calculation time is expected with this technique.

4.3 Ambiguity Resolution

Once an ambiguity is determined from the parsing, it must be resolved, either manually, or automatically by the PAAS system. PAAS uses a municipality name file (this file covers the whole country with around 6000 names and has as its source in the Postal Code Directory tape from Canada Post) in an attempt to resolve an ambiguity.

This methodology is limited to the problems related to municipality names. This is not so bad since these problems account for a good portion of the ambiguous situations, and are easy to detect and to resolve (they don't involve a large amount of data). Future work could examine the usefulness of detecting and resolving more situations.

Finally, no matter how good the software becomes, the unresolvable and the non-existent addresses will remain a problem and should be followed-up manually.

5. CONCLUSION

The results of postal address analysis as accomplished by PAAS are encouraging. It decodes a vast majority of addresses, outputs a very informative code for every component, standardizes and generates an ASK properly, and handles ambiguities. Also, PAAS integrates utilities and interfaces for users and maintainers.

Users have access to an interface which processes their addresses through the four basic functions as well as a facility that handles the addresses in error (on-line processing). A file processor program is also provided.

Also integrated into PAAS is a quality assurance tool for PAAS maintainers. PAAS will evolve in the future with the discoveries of new addresses and obsolete addresses. Making sure that the changes to the system are applied properly is tricky. This maintenance tool ensures that a change to the software doesn't jeopardize any valid addresses properly analyzed in previous versions of the system.

ACKNOWLEDGEMENTS

The author would like to thank J.P. Lozano and M. Vriends who worked on the implementation of PAAS, and B.E. Hill and M. Elsaesser for their help in writing this paper. Acknowledgement is also due to J. Armstrong who experimented with PAAS as well as providing comments on an earlier version of this paper.

REFERENCES

- BARRETT, WILLIAM A., BATES, RODNEY M., GUSTAFSON, DAVID A., and COUCH, JOHN D. (1986). *Compiler Construction*. Science Research Associates Inc.
- CANADA POST CORPORATION (1986). *Postal Codes Directory: Atlantic, Quebec, Ontario and Western regions*.
- DEGUIRE, Y. (1987). Research into the parsing and standardization of free format addresses at Statistics Canada. Internal report, Statistics Canada.
- DREW, J. DOUGLAS, ARMSTRONG, JOHN, VAN BAAREN, ALEX, and DEGUIRE, YVES, (1987). Methodology for construction of address registers using several administrative sources. International Symposium on Statistical Uses of Administrative Data, Ottawa.
- HILL, TED (1986). *MPL A Translator Writing System*. System Documentation, 1-4. Statistics Canada.
- LOZANO, J.P. (1987). Postal Address Analysis System Study. Internal Report, Statistics Canada.
- STATISTICS CANADA (1986). Record Linkage Software User Guide. System documentation, Research and General Systems.
- STATISTICS CANADA (1988). Postal address analysis system (PAAS): Project charter (draft). Internal report. Research and General Systems.

A Brief Note on SQL

DAVID N. EMERY¹

ABSTRACT

This note portrays SQL, highlighting its strengths and weaknesses.

KEY WORDS: Relational database management system; Database query language.

1. INTRODUCTION

A great deal of media attention has been focused on relational database management systems and SQL (pronounced see-quel), the most popular of the associated database query languages. To a large extent, SQL has been cast in the role of panacea for all the ills associated with data management. Unfortunately, this leads to a great deal of misconception on the part of potential users of SQL. These people are then sometimes disappointed with SQL when they eventually get a chance to use it.

The intent of this note is to clear up some of this misconception by providing a realistic portrayal of SQL, highlighting its inherent strengths and weaknesses. No attempt will be made to elaborate the advantages of the relational data model itself. These advantages have been adequately documented elsewhere (Date 1985).

2. SQL - WHAT IS IT?

The interaction which takes place between a user (whether systems developer or end user) and a database management system can be broadly categorized according to the function taking place:

- data definition;
- data control (*i.e.* authorization and control of data integrity);
- data retrieval; and,
- data modification (*i.e.* insert, update, and delete).

A database management system must provide interfaces for carrying out each of these functions. Depending on the particular system, these interfaces take the form of utilities, query languages, and/or subroutine libraries for programming languages.

SQL addresses these four functions in a single well-defined, rigidly structured language. SQL is the interface used to communicate, to the database management system, how relations (*i.e.* logical files or tables) are to be subdivided and/or combined to create new relations.

The key to understanding SQL's capabilities is an appreciation of the fact that SQL addresses exactly these four roles - no more and no less. Any other functionality must be supplied by the application which initiates the SQL statement.

Consider the following example. The table, DWELLING, contains information about dwellings such as number of occupants, type of dwelling, where it is located, type of heating,

¹ David N. Emery, Statistics Canada, Research and General Systems Subdivision, Room 2405, Main Building, Tunney's Pasture, Ottawa, Ontario K1A 0T6.

and age of dwelling. In order to impute the type of dwelling, one might want to obtain a set of potential donor dwellings which are in the same geographic area, are the same age, and use the same heating method. The following SQL statement could be issued to obtain a donor set:

```
SELECT DWELLING__ID, TYPE__OF__DWELLING                (Query 1)
FROM DWELLING
WHERE HEATING__TYPE = 'GAS' AND
      AGE = 20 AND
      LOCATION__CODE = 'XYZ';
```

SQL does not provide a mechanism for manipulating the set of retrieved donor records. Selecting the *n*'th record, every second record, or a random record are all beyond the capability of SQL. Similarly, SQL has no mechanism for manipulating a table to affect its appearance on a terminal or printer. These are capabilities one would rightfully demand of a programming language, and hence the term database query language. Calling SQL a fourth generation language (4GL), then comparing it to products which incorporate only the data retrieval and data modification functions into a programming language, only adds to the confusion. It is really an apples and oranges comparison since both are 4GLs, but of very different flavours.

Given this very focused functionality, the obvious question then has to be — why all the fuss about SQL?

3. SQL — ITS BENEFITS

3.1 Implementation Transparency

A SQL query indicates nothing about how the data is actually organized and stored on the database. The query states what is to be retrieved, modified, or stored; the database management system determines the best way to do it. Issues such as:

- which data columns are indexed (a performance improvement feature);
- whether the table/column is actually stored or merely an execution time combination of other tables; and,
- the data's internal representation (*i.e.* floating point, packed decimal, binary)

have no bearing whatsoever on a SQL statement's syntax. Consequently, the user is immune to changes in the database's organization and structure. Changes to the underlying structure of the database can be made at will without changing the query. A query is immediately able to take advantage of improvements in the database structure or optimization algorithms.

Similarly, when formulating a SQL query the user does not specify the order in which processing is to take place to satisfy the query. That is the responsibility of the query processing software's optimization algorithms. This software evaluates the query against the current structure and organization of the database to determine the most efficient way of satisfying it.

3.2 Non-proprietary, Internationally Accepted Standard

Both the International Standards Organization (ISO) and the American National Standards Institute (ANSI) have recently adopted a common standard for SQL (ISO 1987). The existence of this standard, with a commitment to it by a number of relational database management system vendors, gives software developers access to a much broader market without significantly extra development effort. By building their applications on top of standard SQL, they have removed their reliance on a particular database management system. As a result, the creation

of software tools, built upon an interface to this standard version of SQL, has become a major growth industry. For example, natural language interfaces, fourth generation programming languages, data dictionary software, data entry/validation packages, and spreadsheet software, all layered on top of ANSI/ISO SQL, are beginning to appear on the market.

The active interest in SQL has also had a very positive impact on the SQL standard itself; it is continuing to evolve. The most recent draft revision to the ISO Standard for SQL incorporates the specification of referential integrity constraints into SQL's data definition statements. The significance of this extension to SQL is best illustrated by a further elaboration of the DWELLING example. Assume that the database also has a table PERSONS which contains detailed information about individuals including a dwelling code which indicates the dwelling where they currently reside. One might define a integrity constraint stipulating that each person must be associated with exactly one dwelling. Consequently, it would be an error to delete a DWELLING record which still had any PERSONS records referencing it, or to add a PERSONS record which referenced a nonexistent DWELLING record. Currently, logic to detect and prevent these inconsistencies must be inserted into each application program capable of deleting a DWELLING record. With the incorporation of referential integrity specifications into SQL, this program logic will no longer be required. The DBMS software assumes responsibility for detecting and terminating any attempt to remove a DWELLING record which still has associated PERSONS records.

3.3 Ease of Extension

One of the major differences between the various vendors' versions of SQL is the number and variety of supported functions. This is to a large extent due to the ease with which extra functionality can be incorporated into SQL, without change to its overall structure. For example, the SQL standard documents the grouping functions of average (AVG), maximum (MAX), minimum (MIN), enumeration (COUNT) and aggregation (SUM) for unweighted data. Referring again to the earlier DWELLING example, one could generate various summary statistics about number of occupants, broken down by geographic location:

```
SELECT AVG (NO__OF__OCCUPANTS), MAX (NO__OF__OCCUPANTS), (Query 2)
      MIN (NO__OF__OCCUPANTS), SUM (NO__OF__OCCUPANTS),
      COUNT (NO__OF__OCCUPANTS)
FROM DWELLING
GROUP BY LOCATION__CODE;
```

Some Vendors have augmented these functions with others such as variance (VARIANCE) and standard deviation (STDDEV). With these extra functions the identification of outliers, more than one standard deviation from the mean, is a very straightforward exercise:

```
SELECT DWELLING__ID FROM DWELLING                                     (Query 3)
WHERE NO__OF__OCCUPANTS <
      (SELECT AVG (NO__OF__OCCUPANTS) - STDDEV (NO__OF__OCCUPANTS)
       FROM DWELLING)
OR
      NO__OF__OCCUPANTS >
      (SELECT AVG (NO__OF__OCCUPANTS) + STDDEV (NO__OF__OCCUPANTS)
       FROM DWELLING);
```

3.4 Single Interface to the Database

When interrogating a database from within a host language program such as PL/1, FORTRAN, or C, one also uses SQL statements. These statements are virtually identical to

those used when interrogating the database interactively via a SQL statement processor. The only difference lies in the fact that the host language interface requires an additional INTO clause to indicate the program variables receiving the results of the query.

By using an identical interface to a host programming language, one is able to separate the program development and debugging exercise into two distinct activities:

- testing of the database retrieval storage statements (*i.e.* the SQL statements themselves), and
- testing of the program code which manipulates the data.

The first of these activities can be carried out using a SQL command interpreter even before the host language program has been written. The optimal SQL statements can then be moved directly into the host program where the testing effort can be focused on the logic associated with manipulating the data.

Since the SQL statements embedded in the host language are interpreted at execution time, any changes made to the database organization or structure are immediately reflected in the program.

3.5 Suitability for Distributed Databases/Database Machines

One of the hottest topics in database management systems technology today is distributed databases. In a distributed database environment, the data is spread across a number of different databases (often on physically separate machines). It is the DBMS software's responsibility to intercept a user's query, translate it into appropriate queries to the various constituent databases, and assemble the results of these queries for presentation.

As discussed earlier, a SQL statement is devoid of constructs associated with describing how or where the data is stored on the database. Consequently, in a distributed database environment where SQL is used as the database query language, data can be moved between machines with no change whatsoever to existing applications. SQL is therefore becoming quite popular with the developers of distributed database management systems.

For similar reasons, SQL is gaining popularity as a query language for database machines. These machines take advantage of relational (*i.e.* tabular) data structures' inherent regularity to partition them across a number of parallel processors. These processors have instruction sets specifically designed to perform relational operations. The lack of representational detail in SQL queries completely insulates users from an awareness of what these machines are doing behind the scene.

4. SUMMARY

There is no question that SQL has quickly become the pre-eminent database query language. The database management system which does not feature a SQL interface will soon be the exception. An interesting anomaly will however emerge. The user will, over time, see less and less of SQL. Rather than trying to make SQL itself a user-friendly language, effort will be focused on the development of application specific tools which provide the user with an interface tailored to the task at hand. SQL will be the common interface between these tools and the various databases.

REFERENCES

- DATE, C.J. (1985). *An Introduction to Database Systems*. Don Mills: Addison-Wesley.
- INTERNATIONAL STANDARDS ORGANIZATION (1987). Database Language SQL. International Standards Organization 9075.

A Bibliography on Randomized Response: 1965 - 1987

GAD NATHAN¹

ABSTRACT

A comprehensive bibliography of books, research reports and published papers, dealing with the theory, application and development of randomized response techniques, includes a subject classification.

KEY WORDS: Survey; Sensitive issues; Confidentiality.

1. INTRODUCTION

The recent increase in requirements for extensive data on sensitive issues, (such as the detailed information on sexual behavior, necessary to study the spread of the AIDS epidemic), has lead to renewed examination of the techniques available for obtaining answers to sensitive questions. The difficulties of applying conventional survey techniques to obtain data on sensitive issues in a large-scale survey are well known and several alternative techniques have been proposed - Bradburn and Sudman (1979). The most prominent of these has been the randomized response technique, originally proposed by Warner (1965). The underlying idea is that the respondent uses a random mechanism to select the question to which he answers and the interviewer knows only the response itself, without knowing which question is being answered. This is supposed to reduce biases due to non-response and to response error, by assuring the respondent that his privacy is protected by the method (in that the question he is being asked is unknown to the interviewer) and thereby convincing him to cooperate more readily and to answer more truthfully than he might by a direct question.

Since 1965 a great deal of research into various aspects of the technique has been carried out. This includes theoretical developments, development of new randomization techniques and extensions to quantitative variables, to polytomous questions and to the multivariate case. Problems of estimation, optimization of design parameters and sample design, specific to randomized response, have also been dealt with. A large number of empirical studies using randomized response have been carried out in various application areas, such as studies of drug use, abortions, drunken driving and crime, many of them with some evaluation, often by validation studies. The experience in these studies is very divergent, with some showing marked gains due to the use of randomized response and others showing no gain at all in response rates or in response reliability. Respondents' attitudes to randomized response, their comprehension of the procedure, their perceptions of confidentiality and of the protection that the procedure provides have also been investigated, in attempts to understand the reasons for the differences in the empirical results.

This large body of research is scattered among over 250 theses, research reports, published papers and books, which have appeared, (in at least seven languages), over the last 20 odd years. These include many expository and survey papers and two bibliographies - Kim and

¹Gad Nathan, Department of Statistics, Hebrew University, Mt. Scopus, 91905 Jerusalem, Israel. This bibliography was prepared while serving as Service Fellow at the U.S. National Center for Health Statistics, 3700 East-West Highway, Hyattsville, MD 20782.

Flueck (1976) and Daniel (1979) – the latter an annotated one. Three comprehensive books on the subject – Defaa (1982), Fox and Tracy (1986) and Chaudhuri and Mukerjee (1988) – have also appeared. Unfortunately none of these include a fully comprehensive and updated bibliography and the present one is an attempt to correct this lacuna.

Although an attempt has been made to be as comprehensive as possible, by including both published and unpublished papers, the latter are obviously covered only in as far as information about them was available from various sources. In addition, an attempt was made to reduce duplication by excluding unpublished reports or papers presented at meetings whose content is substantially included in a subsequently published paper. However, Ph.D. theses are generally included, since they usually have more detail than the papers derived from them. Papers about other survey methods for dealing with sensitive issues, which can be considered as alternatives to randomized response, are included only if they relate to a comparison of the alternative to randomized response. Papers dealing with randomization techniques to ensure confidentiality of data already collected (such as random rounding or encoding) are not included, unless they also relate to the use of randomization in the collection process itself.

The bibliography is arranged as an alphabetical listing, which gives full citation details in the standard way used for reference lists. Titles are given in the language of the paper or book, if known. Otherwise, for publications not in English, the title is given in English with a designation of the original language in parentheses. Most of the non-English papers include a summary or abstract in English. A series of letter codes on the right edge of the page, opposite each reference, indicates a classification by subject. The classification categories and codes are given below. An author index and a classified listing by subject, not included due to space limitations, are available from the author.

2. SUBJECT CLASSIFICATION CODES

- A – Applications and field experiments.
- B – Bibliographies and survey papers.
- C – Confidentiality, respondent comprehension, attitude and protection.
- E – Evaluation of alternative techniques or estimators.
- H – Hypothesis testing, estimation and analysis.
- M – Multivariate case.
- O – Optimization of design parameters.
- P – Polytomous questions.
- Q – Quantitative variables.
- R – Randomization devices and techniques.
- S – Sample design.
- T – Theoretical developments.
- V – Validation studies.
- X – Expository papers.

ACKNOWLEDGEMENTS

The author would like to thank the many colleagues and authors of papers on randomized response who reviewed an earlier draft of this bibliography and provided comments, additional references and reprints of their papers. The preparation of this bibliography was partly supported by a National Science Foundation grant (No. SES-8612320).

REFERENCES

ABERNATHY, J.R., GREENBERG, B.G., and HORVITZ, D.G. (1970). Estimates of induced abortion in urban North Carolina. *Demography*, 7, 19-29. AC

ABUL-ELA, A.A. (1966). Randomized response models for sample surveys on human populations. Ph.D. thesis, University of North Carolina, Chapel Hill. APT

ABUL-ELA, A.A., and ABDEL-HAMIED, S.M. (1984). Randomized response ratio estimates: bias and efficiency. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 794-799. EH

ABUL-ELA, A.A., and ABDEL-HAMIED, S.M. (1985). A randomized response ratio estimate from quantitative data. *Proceedings of the Social Statistics Section, American Statistical Association*, 300-305. HQ

ABUL-ELA, A.A., and DAKROURI, H.M. (1980). Randomized response models: a ratio estimator. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 205-208. H

ABUL-ELA, A.A., GREENBERG, B.G., and HORVITZ, D.G. (1967). A multi-proportions randomized response model. *Journal of the American Statistical Association*, 62, 990-1008. P

ADHIKARI, A.K. (1982). On randomized response surveys with sensitive quantitative characters: a case study in the Indian Statistical Institute. Technical Report ASC826, Indian Statistical Institute, Calcutta. A

ADHIKARI, A.K., CHAUDHURI, A., and VIJAYAN, K. (1984). Optimum sampling strategies for randomized response trials (with discussion). *International Statistical Review*, 52, 115-125. BHOT

AHSANULLAH, M., and EICHHORN, B.H. (1984). On scrambled response of sensitive quantitative data. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 800-802. QR

ALBERS, W. (1982). Simple randomized response procedures with bounded respondent risk for quantitative data. *Kwantitatieve Methoden*, 8, 35-46. CQ

ALEXANDER, J.R. (1978). Probability as an aid in social research: the randomized response technique. *Mathematical Spectrum*, 11, 10-13. X

ANDERSON, H. (1975). Efficiency versus protection in randomized response designs. Ph.D. thesis, University of Lund, Sweden. CPT

ANDERSON, H. (1976). Estimation of a proportion through randomized response (with discussion). *International Statistical Review*, 44, 213-217. CQT

ANDERSON, H. (1977). Efficiency versus protection in a general randomized response model. *Scandinavian Journal of Statistics*, 4, 11-19. CQT

BARKSDALE, W.B. (1971). New randomized response techniques for control of non-sampling errors in surveys. Ph.D. thesis, University of North Carolina, Chapel Hill. MT

BARKSDALE, W.B. (1975). New randomized response techniques for control of non-sampling errors in surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 302-304. EMO

BARTH, J.T., and SANDLER, H.M. (1976). Evaluation of the randomized response technique in a drinking survey. *Journal of Studies in Alcoholism*, 37, 690-693. AEV

BASULTO, J. (1982). The randomized response design of Warner: a superpopulation model (In Spanish). *Estadística Española*, 96, 51-61. HT

- BEGIN, G., BOIVIN, M., and BELLERSE, J. (1979). Sensitive data collection through the random response technique: some improvements. *Journal of Psychology*, 101, 53-65. R
- BELDT, S.F., DANIEL, W.W., and GARCHA, B.S. (1982). The Takahasi-Sakasegawa randomized response technique. A field test. *Sociological Methods and Research*, 11, 101-111. AERV
- BELLHOUSE, D.R. (1980). Linear models for randomized response designs. *Journal of the American Statistical Association*, 75, 1001-1004. HOQT
- BERMAN, J., McCOMBS, H., and BORUCH, R.F. (1977). Notes on the contamination method. *Sociological Methods and Research*, 6, 45-62. CR
- BLANGIARDO, G.C. (1978). I campioni in blocco con risposta casualizzata. *Rivista di Statistica Applicata*, 11, 89-96. S
- BLANGIARDO, G.C. (1979). La stratificazione nei campioni in blocco con risposta casualizzata. *Rivista di Statistica Applicata*, 12, 26-36. S
- BLOMQUIST, N., and ERIKSSON, S.A. (1974). A general theory of randomized interviews. Research Report 1974:4, Department of Statistics, University of Gothenburg. MT
- BORUCH, R.F. (1971a). Assuring confidentiality of responses in social research: a note on strategies. *American Sociologist*, 6, 308-311. C
- BORUCH, R.F. (1971b). Maintaining confidentiality of data in educational research: a systematic analysis. *American Psychologist*, 26, 413-430. C
- BORUCH, R.F. (1972). Relations among statistical methods for assuring confidentiality of social research data. *Social Science Research*, 1, 403-414. CE
- BORUCH, R.F. (1982). Methods for resolving privacy problems in social research. In *Ethical Issues in Social Science Research*, (Eds. R.R. Faden, R.J. Wallace, and L. Walters), Baltimore: Johns Hopkins University Press. CX
- BORUCH, R.F., and CECIL, J.S. (1979). *Assuring the confidentiality of social research data*. Philadelphia: University of Pennsylvania Press. CX
- BOURKE, P.D. (1974a). Multi-proportions randomized response using the unrelated question technique. Report No. 74, Errors in Surveys Research Project, Institute of Statistics, University of Stockholm. PR
- BOURKE, P.D. (1974b). Symmetry of response in randomized response designs. Report No. 75, Errors in Surveys Research Project, Institute of Statistics, University of Stockholm. RT
- BOURKE, P.D. (1974c). Vector response in randomized response designs. Report No. 76, Errors in Surveys Research Project, Institute of Statistics, University of Stockholm. T
- BOURKE, P.D. (1975). Randomized response designs for multivariate estimation. Report No. 6, Confidentiality in Surveys Research Project, Institute of Statistics, University of Stockholm. HM
- BOURKE, P.D. (1978). Randomized response designs with symmetric response for multi-proportions estimation. *Statistical Review*, 16, 197-204. MR
- BOURKE, P.D. (1981). On the analysis of some multivariate randomized response designs for categorical data. *Journal of Statistical Planning and Inference*, 5, 165-170. EMR
- BOURKE, P.D. (1982). Randomized response multivariate designs for categorical data. *Communications in Statistics*, Ser. A, 11, 2889-2901. MT
- BOURKE, P.D. (1983). Randomized response designs with attribute-based randomization. *Statistical Review*, 5, 125-132. MR
- BOURKE, P.D. (1984). Estimation of proportions using symmetric randomized response designs. *Psychological Bulletin*, 96, 166-172. MR

- BOURKE, P.D., and DALENIUS, T. (1973). Multi-proportions randomized response using a single sample. Report No. 68, Errors in Surveys Research Project, Institute of Statistics, University of Stockholm. P
- BOURKE, P.D., and DALENIUS, T. (1974a). A note on inadmissible estimates in randomized enquiries. Report No. 72, Errors in Surveys Research Project, Institute of Statistics, University of Stockholm. HT
- BOURKE, P.D., and DALENIUS, T. (1974b). Randomized response models with lying. Report No. 71, Errors in Surveys Research Project, Institute of Statistics, University of Stockholm. HOT
- BOURKE, P.D., and DALENIUS, T. (1976). Some new ideas in the realm of randomized inquiries (with discussion). *International Statistical Review*, 44, 219-221. CMR
- BOURKE, P.D., and MORAN, M.A. (1984). Application of the EM algorithm to randomized response data. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 788-793. HQ
- BOURKE, P.D., and MORAN, M.A. (1986). An alternative EM formulation for randomized response data. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 444-447. HP
- BRADBURN, N., and SUDMAN, S. (1979). Improving interview method and questionnaire design. San-Francisco: Jossey-Bass, 1-13. ACE
- BRESSAN, F. (1983). Warner's scheme of sampling with randomized responses with memory (In Italian). *Rivista di Statistica Applicata*, 16, 85-98. B
- BREWER, K.R.W. (1981). Estimating marihuana usage using randomized response - some paradoxical findings. *Australian Journal of Statistics*, 23, 139-148. ACEMV
- BROWN, G.H. (1975). Randomized inquiry vs. conventional questionnaire method in estimating drug usage rates through mail surveys. Technical Report 75-14, Human Resources Research Organization, Alexandria, Virginia. AE
- BROWN, G.H., and HARDING, F.D. (1973). A comparison of methods of studying illicit drug usage. Technical Report 73-9, Human Resources Research Organization, Alexandria, Virginia. AE
- BUCHMAN, T.A., and TRACY, J.A. (1982). Obtaining responses to sensitive questions: conventional questionnaire versus randomized response technique. *Journal of Accounting Research*, 20, 263-271. AEV
- CAMPBELL, A.A. (1987). Randomized response technique. *Science*, 236, 1049. X
- CAMPBELL, C., and JOINER, B.L. (1973). How to get the answer without being sure you've asked the question. *American Statistician*, 27, 229-231. X
- CARR, J.W., MARASCUILO, L.A., and BUSK, P. (1982). Optimal randomized response models and methods for hypothesis testing. *Journal of Educational Statistics*, 7, 295-310. HO
- CHAUDHURI, A. (1983). Randomized response technique to determine input in crop estimation. *Calcutta Statistical Association Bulletin*, 32, 208-210. AH
- CHAUDHURI, A. (1987). Randomized response surveys of finite populations: a unified approach with quantitative data. *Journal of Statistical Planning and Inference*, 15, 157-165. OQS
- CHAUDHURI, A., and ADHIKARI, A.K. (1981). Sampling strategies with randomized response trials and their properties and relative efficiencies. Technical Report ASC815, Indian Statistical Institute, Calcutta. HT
- CHAUDHURI, A., and MUKERJEE, R. (1985). Optionally randomized response techniques. *Calcutta Statistical Association Bulletin*, 34, 225-229. OR

- CHAUDHURI, A., and MUKERJEE, R. (1987). Randomized response techniques: a review. *Statistica Neerlandica*, 41, 27-44. X
- CHAUDHURI, A., and MUKERJEE, R. (1988). *Randomized response: theory and techniques*. New York: Marcel Dekker. BX
- CHEN, E., CHOW, L.P., and LIU, P.T. (1974). Field studies on the new randomized response techniques. Department of Population Dynamics, Johns Hopkins University, Baltimore. A
- CHEN, T.T. (1978). Log-linear models for the categorical data obtained from randomized response techniques. *Proceedings of the Social Statistics Section, American Statistical Association*, 284-288. H
- CHEN, T.T. (1979). Analysis of randomized response as purposively misclassified data. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 158-163. H
- CHI, I.C., CHOW, L.P., and RIDER, R.V. (1972). The randomized response techniques as used in the Taiwan outcome and pregnancy study. *Studies in Family Planning*, 3, 265-269. AERV
- CHOW, L.P., and LIU, P.T. (1973). A new randomized response technique: the multiple answer model. Department of Population Dynamics, Johns Hopkins University, Baltimore. PR
- CHOW, L.P., GRUHN, W., and CHANG, W.P. (1979). Feasibility of the randomized response technique in rural Ethiopia. *American Journal of Public Health*, 69, 273-276. ARV
- CLICKNER, R.P., and IGLEWICZ, B. (1980). Warner's randomized response technique: the two sensitive questions case. *South African Statistical Journal*, 14, 77-86. HMRS
- COHEN, J.E. (1987). Sexual behavior and randomized responses. *Science*, 236, 1503. BX
- COMSTOCK, G.W., CONDE, J.G., and HELSING, K.J. (1985). A simple randomized response device. *American Journal of Epidemiology*, 122, 187-190. RX
- CURLETTE, W.C. (1980). The randomized response technique: using probability theory to ask sensitive questions. *Mathematics Teacher*, 73, 618-621. X
- DALENIUS, T. (1977). Privacy transformations for statistical information systems. *Journal of Statistical Planning and Inference*, 1, 73-86. C
- DALENIUS, T. (1983). Randomized response. In *Solutions to Ethical and Legal Problems in Social Research*, New York: Academic Press, 237-248. CX
- DALENIUS, T., and VITALE, R.A. (1979). A new randomized response technique for estimating the mean of a distribution. In *Contributions to Statistics, Jaroslav Hajek Memorial Volume*, (Ed. J. Jurechkova), Dordrecht: D. Reidel, 43-47. QR
- DANERMARK, B., and SWENSSON, B. (1987). Measuring drug use among Swedish adolescents. *Journal of Official Statistics*, 3, 439-448. ACE
- DANIEL, W.W. (1979). *Collecting sensitive data by randomized response: an annotated bibliography*. Research Monograph No. 85, College of Business Administration, Georgia State University. B
- DAWES, R.M., and MOORE, M. (1980). Guttman scaling orthodox and randomized responses (In German). In *Attitude Measurement*, (Ed. F. Peterman), Göttingen: Verlag für Psychologie, 117-133. CE
- DAWES, R.M., and SMITH, T.L. (1985). Attitude and opinion measurement. In *Handbook of Social Psychology*, (3rd ed.), (Eds. G. Lindzey, and E. Aronson), Hillsdale: Erlbaum, 509-566. CX
- DEFAA, W. (1982). *Anonymisierte Befragungen mit zufallsverschlüsselten Antworten: Die Randomized-Response-Technik (RRT)*. Frankfurt am Main: Verlag Peter Lang. BX

- DELACEY, P.W. (1975). Randomized conditional response. *Proceedings of the Social Statistics Section, American Statistical Association*, 383-386. HT
- DEVORE, J.L. (1977). A note on the randomized response technique. *Communications in Statistics*, Ser. A, 6, 1525-1534. T
- DEVORE, J.L. (1979). Estimating a population proportion using randomized responses. *Mathematics Magazine*, 52, 38-40. X
- DOWLING, T.A., and SHACHTMAN, R.H. (1975). On the relative efficiency of randomized response models. *Journal of the American Statistical Association*, 70, 84-87. EHO
- DOWNS, T., GILILAND, D.C., and KATZ, L. (1978). Probability in a contested election. *American Statistician*, 32, 122-125. H
- DRAGO, E. (1981). Estimate of the mean and the second moment of a population through randomized response sampling (In Italian). *Rivista di Matematica per le Scienze Economiche e Sociali*, 4, 49-58. EHQ
- DRANE, W. (1975). Randomized response to more than one question. *Proceedings of the Social Statistics Section, American Statistical Association*, 395-397. HM
- DRANE, W. (1976). On the theory of randomized responses to two sensitive questions. *Communications in Statistics*, Ser. A, 5, 565-574; Corrigenda (1976), 5, 1552. HM
- DUFFY, J.C., and WATERTON, J.J. (1984). Randomized response models for estimating the distribution function of a quantitative character. *International Statistical Review*, 52, 165-172. HQ
- DURHAM, A.M., and LICHTENSTEIN, M.J. (1983). Response bias in self-report surveys: evaluation of randomized responses. In *Measurement Issues in Criminal Justice*, (Ed. G.P. Waldo), Beverly Hills: Sage Publications. E
- EDGEELL, E. (1980). Additive constants model: a randomized response technique for eliminating evasiveness to quantitative response. *Psychological Bulletin*, 87, 304-308. Q
- EDGEELL, S.E., HIMMELFARB, S., and CIRA, D.J. (1986). Statistical efficiency of using two quantitative randomized response techniques to estimate correlation. *Psychological Bulletin*, 100, 251-256. EHMQ
- EDGEELL, S.E., HIMMELFARB, S., and DUCHAN, K.L. (1982). Validity of forced responses in a randomized response model. *Sociological Methods and Research*, 11, 89-100. AV
- EICHHORN, B., and HAYRE, L.S. (1983). Scrambled randomized response methods for obtaining sensitive quantitative data. *Journal of Statistical Planning and Inference*, 7, 307-316. EHQR
- ELLEM, D., and HOWSON, A. (1979). Randomized response: a survey technique for sensitive questions. In *Interactive Statistics*, (Ed. D. McNeil), New-York: Elsevier North-Holland, 193-207. X
- EMRICH, L. (1983). Randomized response techniques. In *Incomplete Data in Sample Surveys*, Vol. 2, (Eds. W.G. Madow, I. Olkin and D.B. Rubin), New York: Academic Press, 73-80. B
- ERIKSSON, S.A. (1973a). A new model for randomized response. *International Statistical Review*, 41, 101-113. EHQR
- ERIKSSON, S.A. (1973b). Randomized interviews for sensitive questions. Ph.D. thesis, University of Gothenburg, Sweden. QR
- ERIKSSON, S.A. (1976a). Some sampling theory for surveys with randomized response interviews. Report No. 8, Confidentiality in Surveys Research Project, Institute of Statistics, University of Stockholm. C

- ERIKSSON, S.A. (1976b). Regression analysis of data from randomized interviews. Report No. 17, Confidentiality in Surveys Research Project, Institute of Statistics, University of Stockholm. H
- FERRARI, P. (1978). Il test sequenziale di Wald nel campionamento con risposte casualizzate. *Statistica*, 38, 481-492. HRT
- FERRARI, P. (1984). Two-stage sampling with randomized response and unknown strata sizes (In Italian). *Quaderni di Statistica e Matematica Applicata*, 5, 5-19. S
- FIDLER, D.S., and KLEINKNECHT, R.E. (1977). Randomized response versus direct questioning: two data collection methods for sensitive information. *Psychological Bulletin*, 84, 1045-1049. AE
- FIERING, M.B., and HOOPER, M. (1985). Analysis of disclosure avoidance procedures. *Civil Engineering Systems*, 2, 12-19. MX
- FLIGNER, M.A., POLICELLO, G.E., and SINGH, J. (1977). A comparison of two randomized response survey methods with consideration for the level of respondent protection. *Communications in Statistics*, Ser. A, 6, 1511-1524. CEHR
- FLUECK, J.A., and KIM, J.J. (1976). *Bibliography for randomized response*. Mimeo Series No. 33, Department of Statistics, Temple University. B
- FOLSOM, R.E. (1974). A randomized response validation study: comparison of direct and randomized reporting of DUI arrests. Final Report 25U-807, Research Triangle Institute, Research Triangle Park, North Carolina. AV
- FOLSOM, R.E., GREENBERG, B.G., HORVITZ, D.G., and ABERNATHY, J.R. (1973). The two alternate questions randomized response model for human surveys. *Journal of the American Statistical Association*, 68, 525-530. AEHR
- FOX, J.A., and TRACY, P.E. (1980a). A field-validation of a quantitative randomized response model (with discussion). *Proceedings of the Survey Research Methods Section, American Statistical Association*, 299-304. AHOV
- FOX, J.A., and TRACY, P.E. (1980b). The randomized response approach: applicability to criminal justice research and evaluation. *Evaluation Review*, 4, 601-622. A
- FOX, J.A., and TRACY, P.E. (1984). Measuring associations with randomized response. *Social Science Research*, 13, 188-197. H
- FOX, J.A., and TRACY, P.E. (1986). *Randomized response: a method for sensitive surveys*. Beverly Hills: Sage Publications. BX
- GERSTEL, E.K., BRUCE, J., FOLSOM, R.E., and DURHAM, J. (1974). The effectiveness of the Mecklenburg county alcohol safety action project. Mimeo Report, Research Triangle Institute, Research Triangle Park, North Carolina. AV
- GERSTEL, E.K., MOORE, P., FOLSOM, R.E., and KING, D.A. (1970). Mecklenburg county drinking driving attitude survey. Mimeo Report, Research Triangle Institute, Research Triangle Park, North Carolina. AV
- GEURTS, M.D., ANDRUS, R.R., and REINMUTH, J.E. (1975). Researching shoplifting and other deviant customer behavior using the randomized response design. *Journal of Retailing*, 51, 43-48. A
- GODAMBE, V.P. (1980). Estimation in randomized response trials. *International Statistical Review*, 48, 29-32. HOT
- GOODE, T., and HEINE, W. (1978). Surveying the extent of drug use. *Survey Statistician*, 1, 10-12. A
- GOODSTADT, M.S., and GRUSON, V. (1975). The randomized response technique: a test on drug use. *Journal of the American Statistical Association*, 70, 814-818. A

GOODSTADT, M.S., COOK, G., and GRUSON, V. (1978). The validity of reported drug use: the randomized response technique. *International Journal of Addictions*, 13, 359-367. AE

GOULD, A.L., SHAH, B.V., and ABERNATHY, J.R. (1969). Unrelated question randomized response techniques with two trials per respondent. *Proceedings of the Social Statistics Section, American Statistical Association*, 351-359. HOR

GREENBERG, B.G., ABERNATHY, J.R., and HORVITZ, D.G. (1969). Application of the randomized response technique in obtaining quantitative data. *Proceedings of the Social Statistics Section, American Statistical Association*, 40-43. AQ

GREENBERG, B.G., ABERNATHY, J.R., and HORVITZ, D.G. (1970). A new survey technique and application to the field of public health. *Millbank Memorial Fund Quarterly*, 484, Pt. 2, 39-55. X

GREENBERG, B.G., ABERNATHY, J.R., and HORVITZ, D.G. (1986). Randomized response. In *Encyclopedia of Statistical Sciences (Vol. 7)*, (Eds. S. Kotz, N.L. Johnson and C.B. Read), New York: John Wiley, 540-548. BX

GREENBERG, B.G., ABUL-ELA, A.A., SIMMONS, W.R., and HORVITZ, D.G. (1969). The unrelated question randomized response model: theoretical framework. *Journal of the American Statistical Association*, 64, 520-539. ORST

GREENBERG, B.G., HORVITZ, D.G., and ABERNATHY, J.R. (1974). A comparison of randomized response designs. In *Reliability and Biometry; Statistical Analysis of Lifelength*, (Eds. F. Proschan and R.J. Serfling), Philadelphia: SIAM, 787-815. EHR

GREENBERG, B.G., KUEBLER, R.R., ABERNATHY, J.R., and HORVITZ, D.G. (1971). Application of the randomized response technique in obtaining quantitative data. *Journal of the American Statistical Association*, 66, 243-250. EHQR

GREENBERG, B.G., KUEBLER, R.R., ABERNATHY, J.R., and HORVITZ, D.G. (1977). Respondent hazards in the unrelated question randomized response model. *Journal of Statistical Planning and Inference*, 1, 53-60. CE

GUNEL, E. (1985a). A Bayesian comparison of randomized and voluntary response sampling models. *Communications in Statistics, Ser. A*, 14, 2411-2435. EHT

GUNEL, E. (1985b). On the design of randomized response sampling plan. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 457-459. EHT

HAYASHI, C. (1968). Response errors and biased information. *Annals of the Institute of Statistical Mathematics*, 20, 211-228. CT

HILMAR, N.A. (1968). Anonymity, confidentiality and invasions of privacy: the responsibility of the researcher. *American Journal of Public Health*, 58, 324-330. C

HIMMELFARB, S., and EDGELL, S.E. (1980). Additive constants model: a randomized response technique for eliminating evasiveness to quantitative response questions. *Psychological Bulletin*, 87, 525-530. EQR

HIMMELFARB, S., and EDGELL, S.E. (1982). Note on "The randomized response approach." Addendum to Fox and Tracy. *Evaluation Review*, 6, 279-284. BMQ

HIMMELFARB, S., and LICKTEIG, C. (1982). Social desirability and the randomized response technique. *Journal of Personality and Social Psychology*, 43, 710-717. C

HOCHBERG, Y. (1975). Two stage randomized response schemes for estimating a multinomial. *Communications in Statistics*, 4, 1021-1032. EHPR

HORVITZ, D.G., GREENBERG, B.G., and ABERNATHY, J.R. (1975). Recent developments in randomized response designs. In *A Survey of Statistical Design and Linear Models*, (Ed. J.N. Srivastava), New-York: North Holland, 271-285. B

HORVITZ, D.G., GREENBERG, B.G., and ABERNATHY, J.R. (1976a). Randomized response: a data gathering device for sensitive questions (with discussion). *International Statistical Review*, 44, 181-196. B

- HORVITZ, D.G., GREENBERG, B.G., and ABERNATHY, J.R. (1976b). The randomized response technique. In *Perspectives on Attitude Assessment: Surveys and their Alternatives*, (Eds. H.W. Sinaiko and L.A. Broedling), Champaign: Pendelton Publications. BX
- HORVITZ, D.G., SHAH, B.V., and SIMMONS, W.R. (1967). The unrelated question randomized response model. *Proceedings of the Social Statistics Section, American Statistical Association*, 65-72. AER
- IGLEWITZ, B. (1976). A coding approach to the sensitive question problem. *Proceedings of the Social Statistics Section, American Statistical Association*, 414-415. ER
- IIT Research Institute and the Chicago Crime Commission (1971). A study of organized crime in Chicago. IITRI Project No. H-6031, Report prepared for the Illinois Enforcement Commission, Chicago. A
- KAMMERMANN, L.A., GREENBERG, B.G., and QUADE, D. (1985). Selecting optimal values for π_y in the unrelated question randomized response model. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 470-475. CO
- KIM, J.J. (1978). Randomized response techniques for surveying human populations. Ph.D. thesis, Temple University, Philadelphia. MP
- KIM, J.J. (1987). A further development of randomized response for masking dichotomous variables. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 239-244. MR
- KIM, J.J., and FLUECK, J.A. (1976). A review of randomized response designs and some new results. *Proceedings of the Social Statistics Section, American Statistical Association*, 477-482. BM
- KIM, J.J., and FLUECK, J.A. (1978a). An additive randomized response model. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 351-355. AEMP
- KIM, J.J., and FLUECK, J.A. (1978b). Modifications of the randomized response technique for sampling without replacement. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 346-350. ERS
- KOCH, G.G., ABERNATHY, J.R., and IMREY, P.B. (1975). On a method for study of family size preferences. *Demography*, 12, 57-66. AEH
- KOLATA, G. (1987). How to ask about sex and get honest answers. *Science*, 236, 382. X
- KRAEMER, H.C. (1980). Estimation and testing of bivariate association using data generated by the randomized response technique. *Psychological Bulletin*, 87, 304-308. MP
- KROTKI, K.J., and FOX, B. (1974). The randomized response technique, the interview and the self administered questionnaire: an empirical comparison of fertility reports. *Proceedings of the Social Statistics Section, American Statistical Association*, 367-371. AEV
- KROTKI, K.J., and McDANIEL, S.A. (1975). Three estimates of illegal abortion in Alberta, Canada: Survey, mailback questionnaire and randomized response technique. *Bulletin of the International Statistical Institute*, 46, 67-70. AE
- KROTKI, K.J., and McDANIEL, S.A. (1978). La technique de réponse rendue aléatoire; quelques résultats d'une étude à Edmonton, Alberta. *Population et Famille*, 41, 91-119. A
- LAI, C.D. (1982). *A review of randomized response survey models*. Occasional Publications in Mathematics, 12, Department of Mathematics, Massey University, New Zealand. B
- LAMB, C.W., and STEM, D.E. (1978). An empirical validation of the randomized response technique. *Journal of Marketing Research*, 15, 616-621. AEQV
- LANDENNA, G. (1983). Sampling with randomized responses: a general view (In Italian). *Rivista di Statistica Applicata*, 16, 5-14. B

LANKE, J. (1975). On the choice of the unrelated question in Simmon's version of randomized response. *Journal of the American Statistical Association*, 70, 80-83. EO

LANKE, J. (1976). On the degree of protection in randomized interviews (with discussion). *International Statistical Review*, 44, 197-203. CE

LAVIN, P. (1974). A necessary and sufficient condition for asymptotic masking of the Warner MLE estimate. Report No. 79, Institute of Statistics, University of Stockholm. HT

LERNER, M. (1973). The collection of data on deviant behavior: public policy issues (Abstract). *Bulletin of the International Statistical Institute*, 45, 150. C

LEVY, K.J. (1976a). Reducing the occurrence of omitted or untruthful responses when testing hypotheses concerning proportions. *Psychological Bulletin*, 83, 759-761. HX

LEVY, K.J. (1976b). The randomized response technique and large sample pairwise comparisons among the parameters of k independent binomial populations. *British Journal of Mathematical and Statistical Psychology*, 29, 257-262. EHM

LEVY, K.J. (1977a). The randomized response technique and appropriate sample sizes for selecting the largest value of π from among k binomial populations. *British Journal of Mathematical and Statistical Psychology*, 30, 234-236. EOS

LEVY, K.J. (1977b). The randomized response technique and comparisons among the parameters of k independent binomial populations. *Psychological Bulletin*, 84, 244-246. HM

LEVY, K.J. (1978). Sample size comparisons involving the randomized response technique. *Journal of Experimental Education*, 47, 21-23. ES

LEVY, K.J. (1980). The randomized response technique and large sample tests concerning the parameters of a multinomial distribution. *Educational and Psychological Measurement*, 40, 701-708. EHP

LEYSIEFFER, F.W. (1975). Respondent jeopardy in randomized response procedures. Technical Report M338, Florida State University, Talahassee, Florida. CO

LEYSIEFFER, F.W., and WARNER, S.L. (1976). Respondent jeopardy and optimal designs in randomized response models. *Journal of the American Statistical Association*, 71, 649-656. CEO

LIU, P.T., and CHOW, L.P. (1976a). A new discrete quantitative randomized response model. *Journal of the American Statistical Association*, 71, 72-73. HR

LIU, P.T., and CHOW, L.P. (1976b). The efficiency of the multiple trial randomized response technique. *Biometrics*, 32, 607-618. AER

LIU, P.T., CHEN, C.N., and CHOW, L.P. (1976). A study of the feasibility of Hopkins randomized response models. *Proceedings of the Social Statistics Section, American Statistical Association*, 561-566. AEV

LIU, P.T., CHOW, L.P., and MOSLEY, H.W. (1975). Use of the randomized response technique with a new randomizing device. *Journal of the American Statistical Association*, 70, 329-332. AHR

LOCANDER, W., SUDMAN, S., and BRADBURN, N. (1976). An investigation of interview method, threat and response distortion. *Journal of the American Statistical Association*, 71, 269-275. ACE

LOYNES, R.M. (1976). Asymptotically optimal randomized response procedures. *Journal of the American Statistical Association*, 71, 924-928. OQT

MADIGAN, F.C., ABERNATHY, J.R., HERRIN, A.N., and TAN, C. (1976). Purposive concealment of death in household surveys in Misamis Oriental Province. *Population Studies*, 30, 295-303. AV

MARASINI, D. (1978). La stratificazione nel campionamento con riposte casualizzate. *Statistica*, 38, 493-506. ES

- MARASINI, D. (1981). The randomized response in the two-stage sampling scheme (In Italian). *Quaderni di Statistica e Matematica Applicata*, 4, 81-96. S
- MARASINI, D., and FERRARI, P. (1983). Sampling with randomized responses: estimation and hypotheses testing in case of stratified and two-stage sampling (In Italian). *Rivista di Statistica Applicata*, 16, 15-41. HS
- MARASINI, D., and OLIVIERI, D. (1983). Randomized response models and the quality of statistical data (In Italian). *Societa Italiana di Statistica, Atti del Convegno*, 1, 489-513. X
- MATLOFF, N.S. (1984). Use of covariates in randomized response settings. *Statistics and Probability Letters*, 2, 31-34. HM
- MAZZALI, A. (1983). A scheme of sampling with randomized responses in case of k questions (In Italian). *Rivista di Statistica Applicata*, 16, 99-105. M
- McDANIEL, S.A., and KROTKI, K.J. (1979). Estimates of the rate of illegal abortion and the effect of eliminating therapeutic abortion, Alberta 1973-74. *Canadian Journal of Public Health*, 70, 393-398. A
- MILLER, J.D. (1984). A new survey technique for studying deviant behavior. Ph.D. thesis, George Washington University, Washington, DC. E
- MOORS, J.J.A. (1971). Optimization of the unrelated question randomized response model. *Journal of the American Statistical Association* 66, 627-629. EO
- MOORS, J.J.A. (1981). Inadmissibility of linearly invariant estimators in truncated parameter spaces. *Journal of the American Statistical Association*, 76, 910-915. HT
- MOORS, J.J.A. (1985). Estimation in truncated parameter spaces. Ph.D. thesis, Tilburg University, The Netherlands. H
- MORIARTY, M., and WISEMAN, F. (1976). On the choice of a randomization technique with the randomized response model. *Proceedings of the Social Statistics Section, American Statistical Association*, 624-626. CR
- MUKERJEE, R. (1981). Inference on confidential characters from survey data. *Calcutta Statistical Association Bulletin*, 30, 77-88. C
- MUKHOPADHYAY, P. (1980a). A survey on the socio-economic conditions of some college students of Calcutta. Project Report, Indian Statistical Institute. A
- MUKHOPADHYAY, P. (1980b). On estimation of some confidential characters from survey data. *Calcutta Statistical Association Bulletin*, 29, 133-141. X
- MUKHOPADHYAY, P., and HALDER, A.K. (1980). Bayesian tables for Warner's randomized response probabilities. Technical Report ASC802, Indian Statistical Institute, Calcutta. OT
- OLIVIERI, D. (1981). Le risposte casualizzate: stime, dimensioni ed errori campionari. *Rivista di Statistica Applicata*, 14, 79-98. EH
- OLIVIERI, D. (1982). *La diffusione della droga nelle scuole secondarie superiori di Verona*. Vicenza, Italy: Cassa di Risparmio di Verona Vicenza e Belluno. A
- OLIVIERI, D. (1983a). On a modification of Simmon's scheme of sampling with randomized responses, with efficiency comparisons (In Italian). *Rivista di Statistica Applicata*, 16, 57-75. ER
- OLIVIERI, D. (1983b). Stratified sampling with randomized responses and fixed alternative response (In Italian). *Rivista di Statistica Applicata*, 16, 77-84. S
- OLIVIERI, D. (1984). Estimation of parameters and efficiency in the Poole's randomized response model (In Italian). *Societa Italiana di Statistica, Atti dela Riunione Scientifica*, 32, 463-472. E

- OLIVIERI, D., and BRESSAN, F. (1984). A randomized response model with fixed alternative answer (In Italian). *Rivista di Statistica Applicata*, 17, 165-172. E
- ORWIN, R.G., and BORUCH, R.F. (1982). RRT meets RDD: statistical strategies for assuring response privacy in telephone surveys. *Public Opinion Quarterly*, 46, 560-571. CE
- O'BRIEN, D.M., and COCHRAN, R.S. (1977). The comprehensiveness factor in randomized response. *Proceedings of the Social Statistics Section, American Statistical Association*, 270-272. C
- O'BRIEN, D.M., and COCHRAN, R.S. (1978). The effect of less than complete truthfulness on a quantitative randomized response model. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 743-747. CEQ
- O'HAGAN, A. (1987). Bayes linear estimators for randomized response models. *Journal of the American Statistical Association*, 82, 584-587. HT
- PEARL, R.L., and FEDERER, W.T. (1975). Varying levels of probability for selecting sensitive questions using a randomized response technique. *Proceedings of the Social Statistics Section, American Statistical Association*, 584-587. AE
- PITZ, G.E. (1980). Bayes analysis of random response models. *Psychological Bulletin*, 87, 209-212. T
- POHL, B.B., and POHL, N.F. (1975). Random response techniques for reducing non-sampling error in interview survey research. *Journal of Experimental Education*, 44, 48-53. X
- POLLOCK, K.H., and BEK, Y. (1976). A comparison of three randomized response models for quantitative data. *Journal of the American Statistical Association*, 71, 884-886. EQR
- POOLE, W.K. (1974). Estimation of the distribution function of a continuous type random variable through randomized response. *Journal of the American Statistical Association*, 69, 1002-1005. QR
- POOLE, W.K., and CLAYTON, C.A. (1982). Generalizations of a contamination model for continuous type random variables. *Communications in Statistics, Ser. A*, 11, 1733-1742. MQ
- RAGHAVARAO, D. (1978). On an estimation problem in Warner's randomized response technique. *Biometrics*, 34, 87-90. EH
- RAGHAVARAO, D., and FEDERER, W.T. (1979). Block total response as an alternative to the randomized response method in surveys. *Journal of the Royal Statistical Society, Ser. B*, 41, 40-45. ER
- REASER, J.M., HARTSOCK, S., and HOEHN, A.J. (1975). A test of the forced alternative random response questionnaire technique. Technical Report 75-9, Human Resources Research Organization, Alexandria, VA. AE
- REINMUTH, J.E., and GEURTS, M.D. (1975). The collection of sensitive information using a two-stage randomized response model. *Journal of Marketing Research*, 12, 402-407. AEQ
- ROSENBERG, M.J. (1979). Multivariable analysis of a randomized response technique for statistical disclosure control. Ph.D. thesis, University of Michigan, Ann Arbor. CM
- ROSENBERG, M.J. (1980). Categorical data analysis by a randomized response technique for statistical disclosure control (with discussion). *Proceedings of the Survey Research Methods Section, American Statistical Association*, 311-316. AEHP
- ROSENBERG, M.J. (1985). An application of PROC FUNCAT to randomized response data. *Proceedings of the SAS Users Group International Conference*, 10, 1070-1075. H

- ROSENBLATT, R.R., and KELLY, E.L. (1978). A comparison of the sensitivity of the unrelated question randomized response model with three other data accumulation techniques using examination cheating as a model. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 356-361. AE
- SAKASEGAWA, H., and TAKAHASI, K. (1974). Effects of repetition and finite population corrections in randomized response models (In Japanese). *Proceedings of the Institute of Statistical Mathematics*, 22, 59-67. ER
- SAKASEGAWA, H., TAKAHASI, K., and SUZUKI, T. (1977). An investigation of a new randomized response model (In Japanese). *Proceedings of the Institute of Statistical Mathematics*, 25. RT
- SCHEERS, N.J., and DAYTON, C.M. (1982). The covariate unrelated question randomized response model. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 407-410. AER
- SCHEERS, N.J., and DAYTON, C.M. (1986). RRCOV: Computer program for covariate randomized response models. *American Statistician*, 40, 229. H
- SCHMEIDLER, J. (1987). Assessing quality of randomized response: were instructions followed? *Proceedings of the Survey Research Methods Section, American Statistical Association*, 245-249. AHV
- SEN, P.K. (1974a). On the estimation of basic distributions in randomized response models. *Communications in Statistics*, 3, 1081-1092. HQT
- SEN, P.K. (1974b). On unbiased estimation for randomized response models. *Journal of the American Statistical Association*, 69, 997-1001. EHT
- SEN, P.K. (1976). Asymptotically optimal estimators of general parameters in randomized response models (with discussion). *International Statistical Review*, 44, 223-224. HRT
- SHIMIZU, I.M., and BONHAM, G.S. (1978). Randomized response technique in a national survey. *Journal of the American Statistical Association*, 73, 35-39. AE
- SHOTLAND, R.L., and YANKOVSKI, L.D. (1982). The random response method: a valid and ethical indicator of the "truth" in reactive situations. *Personality and Social Psychology Bulletin*, 8, 174-179. AE
- SILVA, L.C. (1983). The randomized response technique: a general model for polytomous variables (In Spanish). *Investigacion Operacional*, 4, 75-100. P
- SIMMONS, W.R. (1970). Response to randomized inquiries: a technique for reducing bias. *Administrative Applications Division Conference Transaction, American Society for Quality Control*, Chapter 4-B. RT
- SINGH, J. (1976). A note on the randomized response technique. *Proceedings of the Social Statistics Section, American Statistical Association*, 772. EH
- SINGH, J. (1978). A note on maximum likelihood from randomized response models. *Proceedings of the Social Statistics Section, American Statistical Association*, 282-283. EH
- SMITH, E.P., and SOSNOWSKI, T.S. (1972). Faculty evaluations by randomized response sampling. *Journal of Experimental Education*, 41, 70-72. AX
- SMITH, L.L., FEDERER, W.T., and RAGHAVARAO, D. (1974). A comparison of three techniques for eliciting truthful answers to sensitive questions. *Proceedings of the Social Statistics Section, American Statistical Association*, 447-452. AE
- SOEKEN, K.L. (1987). Randomized response methodology in health research. *Evaluation and Health Professions*, 10, 58-66. BX
- SOEKEN, K.L., and MACREADY, G.B. (1982). Respondents' perceived protection when using randomized response. *Psychological Bulletin*, 92, 487-489. AEH

SOEKEN, K.L., and MACREADY, G.B. (1985). Randomized response parameters as factors in frequency estimates. <i>Educational and Psychological Measurement</i> , 45, 89-100.	ACE
SPURRIER, J.D., and PADGETT, W.J. (1980). The application of Bayesian techniques in randomized response. In <i>Sociological Methodology</i> , San-Francisco: Jossey-Bass, 533-544.	HT
STEM, D.E., and LAMB, C.W. (1981). The marble-drop technique: a procedure for gathering sensitive information. <i>Decision Science</i> , 12, 702-707.	AE
STEM, D.E., and STEINHORST, R.K. (1984). Telephone interview and mail questionnaire applications of the randomized response model. <i>Journal of the American Statistical Association</i> , 79, 555-564.	AEMQ
SUZUKI, T., TAKAHASI, K., and SAKASEGAWA, H. (1976). Some notes on randomized response techniques (In Japanese). <i>Proceedings of the Institute of Statistical Mathematics</i> , 24, 1-13.	AE
SWENSSON, B. (1972). Stratified randomized response with the special case: a combined use of regular interview and randomized response interview. Report No. 45, Errors in Surveys Research Project, Institute of Statistics, University of Stockholm.	ES
SWENSSON, B. (1976a). A note on relations among one-sample randomized response techniques for dichotomies. Report No. 12, Confidentiality in Surveys Research Project, Institute of Statistics, University of Stockholm.	CE
SWENSSON, B. (1976b). Combined independent questions versus randomized response, efficiencies under equal degree of protection. Report No. 15, Confidentiality in Surveys Research Project, Institute of Statistics, University of Stockholm.	CE
SWENSSON, B. (1976c). Using mixtures of techniques for estimating sensitive attributes. Report No. 13, Confidentiality in Surveys Research Project, Institute of Statistics, University of Stockholm.	E
SWENSSON, B. (1977). Survey measurement of sensitive attributes. Ph.D. thesis, University of Stockholm.	E
TAKAHASI, K., and SAKASEGAWA, H. (1977). A randomized response technique without making use of any randomizing device. <i>Annals of the Institute of Statistical Mathematics</i> , 29, 1-8.	HR
TAMHANE, A.C. (1977). A randomized response technique for investigating several sensitive attributes. <i>Proceedings of the Social Statistics Section, American Statistical Association</i> , 273-278.	CEM
TAMHANE, A.C. (1981). Randomized response techniques for multiple sensitive attributes. <i>Journal of the American Statistical Association</i> , 76, 916-923.	AEHM
TRACY, P.E., and FOX, J.A. (1981). The validity of randomized response for sensitive measurements. <i>American Sociological Review</i> , 46, 187-200.	AEV
VERDOOREN, L.R. (1976). Loten bij delicate vragen: een overzicht van "randomized response" technieken. <i>Statistica Neerlandica</i> , 30, 7-24.	B
WARNER, S.L. (1965). Randomized response: a survey technique for eliminating evasive answer bias. <i>Journal of the American Statistical Association</i> , 60, 63-69.	EHRT
WARNER, S.L. (1971). The linear randomized response model. <i>Journal of the American Statistical Association</i> , 66, 884-888.	MQRT
WARNER, S.L. (1976). Optimal randomized response models (with discussion). <i>International Statistical Review</i> , 44, 205-212.	CEOT
WINKLER, R.L., and FRANKLIN, L.A. (1979). Warner's randomized response model: a Bayesian approach. <i>Journal of the American Statistical Association</i> , 74, 207-214.	HT

- WISEMAN, F., MORIARTY, M., and SCHAFER, M. (1975). Estimating public opinion with the randomized response model. *Public Opinion Quarterly*, 39, 507-513. X
- ZDEP, S.M., and RHODES, I.N. (1976). Making the randomized response technique work. *Public Opinion Quarterly*, 40, 531-537. AE
- ZDEP, S.M., RHODES, I.N., SCHWARZ, R.M., and KILKENNY, M.J. (1979). The validity of the randomized response technique. *Public Opinion Quarterly*, 43, 544-549. AV

'On the Stratification of Skewed Populations' by P. Lavallée and M.A. Hidioglou, Survey Methodology (1988), 14, 33-43.

Formula (3.10), for the computation of $b''_{(h)}$ should be

$$b''_{(h)} = \frac{-\beta'_h + \sqrt{\beta'^2_h - 4\alpha'_h\gamma'_h}}{2\alpha'_h}, h = 1, \dots, L - 1.$$

Its finite population analogue on page 39, should also be as above.

ACKNOWLEDGEMENTS

The Survey Methodology Journal wishes to thank the following persons who have served as referees between February 1, 1988 and December 31, 1988. An asterisk indicates that the person served more than once.

- M.G. Arellano, *University of California*
- * J. Armstrong, *Statistics Canada*
- T.R. Balakrishnan, *University of Western Ontario*
- * M. Bankier, *Statistics Canada*
- * D. Bellhouse, *University of Western Ontario*
- L. Biggeri, *University of Florence*
- J.-R. Boudreau, *Statistics Canada*
- G. Brackstone, *Statistics Canada*
- R.D. Burgess, *Statistics Canada*
- R. Butcher, *U.K. Office of Population Censuses and Surveys*
- M. Cairns, *Office of Privatization and Regulatory Affairs*
- * R.G. Carter, *Statistics Canada*
- D. Chapman, *U.S. Bureau of the Census*
- * S. Cheung, *Statistics Canada*
- N. Chinnappa, *Statistics Canada*
- G.H. Choudhry, *Statistics Canada*
- C. Courchesne, *Bureau de la Statistique du Québec*
- M.L. Cohen, *University of Maryland*
- F.R. Cronkhite, *U.S. Bureau of Labor Statistics*
- P. Demmons, *Statistics Canada*
- T. Dielman, *Texas Christian University*
- D. Drew, *Statistics Canada*
- J.L. Eltinge, *Iowa State University*
- E.P. Ericksen, *Temple University*
- V.M. Estevao, *Statistics Canada*
- R.E. Fay, *U.S. Bureau of the Census*
- R. Fecso, *U.S. Department of Agriculture*
- I.P. Fellegi, *Statistics Canada*
- M. Fluet, *SOM Inc.*
- R.E. Folsom, *Research Triangle Institute*
- J. Gambino, *Statistics Canada*
- J.F. Gentleman, *Statistics Canada*
- M.-F. Germain, *Statistics Canada*
- M. Gonzalez, *U.S. Office of Management and Budget*
- * G.B. Gray, *Statistics Canada*
- S.G. Heeringa, *University of Michigan*
- * M.A. Hidirolou, *Statistics Canada*
- T. Hill, *Statistics Canada*
- R.R. Hocking, *Texas A & M University*
- H. Hogan, *U.S. Bureau of the Census*
- D. Holt, *University of Southampton*
- T.B. Jabine, *Statistical Consultant*
- J. Kadane, *Carnegie-Mellon University*
- G. Kalton, *University of Michigan*
- B.F. King, *Educational Testing Service*
- B.F. Klugh, Jr., *U.S. Department of Agriculture*
- P. Krishnan, *University of Alberta*
- * S. Kumar, *Statistics Canada*
- N.M. Lalu, *University of Alberta*
- P. Lavallée, *Statistics Canada*
- B. Lefrançois, *Statistics Canada*
- H. Lee, *Statistics Canada*
- J. Leyes, *Statistics Canada*
- R. Little, *University of California*
- L. Mach, *Statistics Canada*
- K. Namboodiri, *Ohio State University*
- J.L. O'Brien, *U.S. Bureau of the Census*
- S.C. Puri, *Agriculture Canada*
- B. Quenneville, *Statistics Canada*
- * J.N.K. Rao, *Carleton University*
- T.J. Rao, *Indian Statistical Institute*
- J.G. Robinson, *U.S. Bureau of the Census*
- * D. Royce, *Statistics Canada*
- I. Sande, *Statistics Canada*
- C. Särndal, *University of Montreal*
- F. Scheuren, *U.S. Internal Revenue Service*
- V.A. Sposito, *Iowa State University*
- T.W.F. Stroud, *Queen's University*
- A. Stuart, *London School of Economics*
- J. Sullivan, *U.S. Bureau of the Census*
- J.-L. Tambay, *Statistics Canada*
- * V. Tremblay, *Statplus*
- K. Wachter, *University of California*
- M.J. Wenzowski, *Statistics Canada*
- G.S. Werking, *U.S. Bureau of Labor Statistics*
- W.E. Winkler, *U.S. Bureau of the Census*

Acknowledgements are also due to those who assisted during the production of the 1987 issues: C. VanBastelaar (Photocomposition), G. Gaulin (Author Services) and M. Haight (Translation Services).

We would like to thank the staff of Social Survey Methods and Business Survey Methods Divisions who assisted in proofreading and verification. Finally we wish to acknowledge J. Clarke, E. Corriveau, J. Dufresne, M. Kent, C. Larabie, D. Lemire and N. Smalldridge for their support with coordination, typing and copy editing.

Applied Statistics

JOURNAL OF THE ROYAL STATISTICAL SOCIETY (SERIES C)

CONTENTS

Volume 38, No. 1, 1989

	<i>Page</i>
Space-time modelling with long-memory dependence: assessing Ireland's wind power resource <i>J. Haslett and A. E. Raftery</i>	1
Case-weighted measures of influence for proportional hazards regression <i>A. N. Pettitt and I. Bin Daud</i>	51
Approximating the normal tail probability and its inverse for use on a pocket calculator <i>J.-T. Lin</i>	69
Unweighted sum of squares test for proportions <i>J. B. Copas</i>	71
On the role of cause-of-death data in the analysis of rodent tumorigenicity experiments <i>L. E. Archer and L. M. Ryan</i>	81
A probabilistic model of squash: strategies and applications <i>J. Simmons</i>	95
A note on 'random' purchasing: additional insights from Dunn, Reader and Wrigley <i>B. E. Kahn and D. G. Morrison</i>	111
The use of guided reformulations when collinearities are present in non-linear regression <i>J. S. Simonoff and C.-L. Tsai</i>	115
Multiple-spell regression models for duration data <i>A. Hamerle</i>	127
Rotation of ill-defined principal components <i>I. T. Jolliffe</i>	139
Construction of row and column designs with contiguous replicates <i>E. R. Williams and J. A. John</i>	149
<i>Statistical Software Reviews</i>	155
<i>Statistical Algorithms</i>	
AS 242 The exact likelihood of a vector autoregressive moving average model <i>B. L. Shea</i>	161
AS 243 Cumulative distribution function of the non-central t distribution <i>R. V. Lenth</i>	185
AS 244 Decomposability and collapsibility for log-linear models <i>Z. Geng</i>	189
<i>Remarks</i>	
AS R76 A remark on Algorithm AS 215: Maximum-likelihood estimation of the parameters of the generalized extreme-value distribution <i>A. J. Macleod</i>	198
AS R77 A remark on Algorithm AS 152: Cumulative hypergeometric probabilities <i>B. L. Shea</i>	199
<i>Correction</i>	
Correction to Algorithm AS 231: The distribution of a noncentral χ^2 variable with nonnegative degrees of freedom <i>R. W. Farebrother</i>	204

Published in three parts per year. Annual subscription £23.00; single issues £10.00. The Algorithm Section is available separately—annual subscription £4.00. All communications should be addressed to the Executive Secretary, Royal Statistical Society, 25 Enford Street, London W1H 2BH, UK.

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 10, No. 2 and onward) of *Survey Methodology* as a guide and note particularly the following points:

1. Layout

- 1.1 Manuscripts should be typed on white bond paper of standard size ($8\frac{1}{2} \times 11$ inch), one side only, entirely double spaced with margins of at least $1\frac{1}{2}$ inches on all sides.
- 1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(\cdot)" and "log(\cdot)", etc.
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (e.g., w , ω ; o , O ; 0 ; l , 1).
- 3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de Techniques d'enquête (à partir du vol. 10, n° 2) et de noter les points suivants:

1. Présentation

- 1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8 1/2 par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1 1/2 pouce tout autour.
- 1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
- 1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
- 1.4 Les remerciements doivent paraître à la fin du texte.
- 1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

2. Résumé

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. Rédaction

- 3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
- 3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme $\exp(-)$ et $\log(-)$ etc.
- 3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
- 3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
- 3.5 Distinguer clairement les caractères ambigus (comme w, ω ; o, O; 1, l).
- 3.6 Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.

4. Figures et tableaux

- 4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
- 4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois.)

5. Bibliographie

- 5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.
- 5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

Page	
1	Space-time modelling with long-memory dependence: assessing Ireland's wind power resource <i>J. Haslett and A. E. Raftery</i>
51	Case-weighted measures of influence for proportional hazards regression <i>A. N. Pettitt and I. Bin Daud</i>
69	Approximating the normal tail probability and its inverse for use on a pocket calculator <i>J.-T. Lin</i>
71	Unweighted sum of squares test for proportions <i>J. B. Copas</i>
81	On the role of cause-of-death data in the analysis of rodent tumorigenicity experiments <i>L. E. Archer and L. M. Ryan</i>
95	A probabilistic model of squash: strategies and applications <i>J. Simmons</i>
111	A note on 'random' purchasing: additional insights from Dunn, Reader and Wrigley <i>B. E. Kahn and D. G. Morrison</i>
115	The use of guided reformulations when collinearities are present in non-linear regression <i>J. S. Simonoff and C.-L. Tsai</i>
127	Multiple-spl regression models for duration data <i>A. Hamerle</i>
139	Rotation of ill-defined principal components <i>I. T. Jolliffe</i>
149	Construction of row and column designs with contiguous replicates <i>E. R. Williams and J. A. John</i>
155	Statistical Software Reviews
161	Statistical Algorithms
185	AS 242 The exact likelihood of a vector autoregressive moving average model <i>B. L. Shea</i>
189	AS 243 Cumulative distribution function of the non-central t distribution <i>R. V. Lenth</i>
198	AS 244 Decomposability and collapsibility for log-linear models <i>Z. Geng</i>
199	Remarks
204	AS R76 A remark on Algorithm AS 215: Maximum-likelihood estimation of the parameters of the generalized extreme-value distribution <i>A. J. Macleod</i>
204	AS R77 A remark on Algorithm AS 152: Cumulative hypergeometric probabilities <i>B. L. Shea</i>
204	Correction to Algorithm AS 231: The distribution of a noncentral χ^2 variable with nonnegative degrees of freedom <i>R. W. Farebrother</i>

Published in three parts per year. Annual subscription £23.00; single issues £10.00. The Algorithm Section is available separately—annual subscription £4.00. All communications should be addressed to the Executive Secretary, Royal Statistical Society, 25 Enford Street, London W1H 2BH, UK.

REMERCIEMENTS

La revue *Techniques d'enquête* désire remercier les personnes suivantes, qui ont accepté de faire la critique d'un article entre le premier février 1988 et le 31 décembre 1988. Une astérisque indique que la personne a participé plus d'une fois.

- M.G. Arellano, *University of California*
 * J. Armstrong, *Statistique Canada*
 T.R. Balakrishnan, *University of Western Ontario*
 * M. Bankier, *Statistique Canada*
 * D. Bellhouse, *University of Western Ontario*
 L. Biggieri, *Université de Florence*
 J.-R. Boudreau, *Statistique Canada*
 G. Brackstone, *Statistique Canada*
 R.D. Burgess, *Statistique Canada*
 R. Bucher, *U.K. Office of Population Censuses and Surveys*
 M. Cairns, *Bureau de Privatisation et Affaires Réglementaires*
 * R.G. Carter, *Statistique Canada*
 D. Chapman, *U.S. Bureau of the Census*
 * S. Cheung, *Statistique Canada*
 N. Chinappa, *Statistique Canada*
 G.H. Choudhry, *Statistique Canada*
 C. Courchesne, *Bureau de la Statistique du Québec*
 M.L. Cohen, *University of Maryland*
 F.R. Cronkhitte, *U.S. Bureau of Labor Statistics*
 P. Demmons, *Statistique Canada*
 T. Dielman, *Texas Christian University*
 D. Drew, *Statistique Canada*
 J.L. Eitinge, *Iowa State University*
 E.P. Ericksen, *Temple University*
 V.M. Estevo, *Statistique Canada*
 R.E. Fay, *U.S. Bureau of the Census*
 R. Fecso, *U.S. Department of Agriculture*
 I.P. Fellegi, *Statistique Canada*
 M. Finet, *SOM Inc.*
 R.E. Folsom, *Research Triangle Institute*
 J. Gambino, *Statistique Canada*
 J.F. Gentleman, *Statistique Canada*
 M.-F. Germain, *Statistique Canada*
 M. Gonzalez, *U.S. Office of Management and Budget*
 * G.B. Gray, *Statistique Canada*
 S.G. Heeringa, *University of Michigan*
 * M.A. Hidiroglou, *Statistique Canada*
 T. Hill, *Statistique Canada*
- R.R. Hocking, *Texas A & M University*
 H. Hogan, *U.S. Bureau of the Census*
 D. Holt, *University of Southampton*
 T.B. Jabine, *Expert-conseil en statistique*
 J. Kadane, *Carnegie-Mellon University*
 G. Kalton, *University of Michigan*
 B.F. King, *Educational Testing Service*
 B.F. Klugh, Jr., *U.S. Department of Agriculture*
 P. Krishnan, *University of Alberta*
 * S. Kumar, *Statistique Canada*
 N.M. Lahu, *University of Alberta*
 P. Lavallée, *Statistique Canada*
 B. Lefrançois, *Statistique Canada*
 H. Lee, *Statistique Canada*
 J. Leyes, *Statistique Canada*
 R. Little, *University of California*
 L. Mach, *Statistique Canada*
 K. Namboodiri, *Ohio State University*
 J.L. O'Brien, *U.S. Bureau of the Census*
 S.C. Puri, *Agriculture Canada*
 B. Quenneville, *Statistique Canada*
 * J.N.K. Rao, *Carleton University*
 T.J. Rao, *Indian Statistical Institute*
 J.G. Robinson, *U.S. Bureau of the Census*
 * D. Royce, *Statistique Canada*
 I. Sande, *Statistique Canada*
 C. Sarnadal, *Université de Montréal*
 F. Scheuren, *U.S. Internal Revenue Service*
 V.A. Sposito, *Iowa State University*
 T.W.F. Stroud, *Queen's University*
 A. Stuart, *London School of Economics*
 J. Sullivan, *U.S. Bureau of the Census*
 J.-L. Tambay, *Statistique Canada*
 * V. Tremblay, *Statplus*
 K. Wachter, *University of California*
 M.J. Wenzowski, *Statistique Canada*
 G.S. Werking, *U.S. Bureau of Labor Statistics*
 W.E. Winkler, *U.S. Bureau of the Census*

On remercie également ceux qui ont contribué à la production des numéros de la revue pour 1988: C. VanBastelaer (Photocomposition), G. Gaulin (Services aux auteurs) and M. Haighi (Services de traduction). On tient à remercier le personnel des Divisions des méthodes d'enquêtes sociales et des méthodes d'enquêtes-entreprises qui ont aidé à la correction et à la vérification. Finalement on désire exprimer notre reconnaissance à J. Clarke, J. Dufréne, M. Kent, C. Larabie et D. Lemire pour leur apport à la coordination, la dactylographie et à la rédaction.

'Sur la stratification de populations asymétriques' par P. Lavallée et M.A. Hidiroglou, Techniques d'enquête (1988), 14, 35-45.

L'équation (3.10), pour le calcul de $b''_{(h)}$ devrait être

$$b''_{(h)} = \frac{-\beta'_h + \sqrt{\beta'^2_h - 4\alpha'_h\gamma'_h}}{2\alpha'_h}, h = 1, \dots, L - 1.$$

Son analogue pour population finie à la page 41 devrait aussi être tel qu'indiqué ci-dessus.

- STEM, D.E., et LAMB, C.W. (1981). The marble-drop technique: a procedure for gathering sensitive information. *Decision Science*, 12, 702-707.
- STEM, D.E., et STEINHORST, R.K. (1984). Telephone interview and mail questionnaire applications of the randomized response model. *Journal of the American Statistical Association*, 79, 555-564.
- SUZUKI, T., TAKAHASI, K., et SAKASEGAWA, H. (1976). Some notes on randomized response techniques (en japonais). *Proceedings of the Institute of Statistical Mathematics*, 24, 1-13.
- SWENSSON, B. (1972). Stratified randomized response with the special case: a combined use of regular interview and randomized response interview. Report No. 45, Errors in Surveys Research Project, Institute of Statistics, Université de Stockholm.
- SWENSSON, B. (1976a). A note on relations among one-sample randomized response techniques for dichotomies. Report No. 12, Confidentiality in Surveys Research Project, Institute of Statistics, Université de Stockholm.
- SWENSSON, B. (1976b). Combined independent questions versus randomized response, efficiencies under equal degree of protection. Report No. 15, Confidentiality in Surveys Research Project, Institute of Statistics, Université de Stockholm.
- SWENSSON, B. (1976c). Using mixtures of techniques for estimating sensitive attributes. Report No. 13, Confidentiality in Surveys Research Project, Institute of Statistics, Université de Stockholm.
- SWENSSON, B. (1977). Survey measurement of sensitive attributes. Thèse de doctorat, Université de Stockholm.
- TAKAHASI, K., et SAKASEGAWA, H. (1977). A randomized response technique without making use of any randomizing device. *Annals of the Institute of Statistical Mathematics*, 29, 1-8.
- TAMAHANE, A.C. (1977). A randomized response technique for investigating several sensitive attributes. *Proceedings of the Social Statistics Section, American Statistical Association*, 273-278.
- TAMAHANE, A.C. (1981). Randomized response techniques for multiple sensitive attributes. *Journal of the American Statistical Association*, 76, 916-923.
- TRACY, P.E., et FOX, J.A. (1981). The validity of randomized response for sensitive measurements. *American Sociological Review*, 46, 187-200.
- VERDOOREN, L.R. (1976). Loten bij delicate vragen: een overzicht van "randomized response" technieken. *Statistica Neerlandica*, 30, 7-24.
- WARNER, S.L. (1965). Randomized response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60, 63-69.
- WARNER, S.L. (1971). The linear randomized response model. *Journal of the American Statistical Association*, 66, 884-888.
- WARNER, S.L. (1976). Optimal randomized response models (with discussion). *International Statistical Review*, 44, 205-212.
- WINKLER, R.L., et FRANKLIN, L.A. (1979). Warner's randomized response model: a Bayesian approach. *Journal of the American Statistical Association*, 74, 207-214.
- WISEMAN, F., MORIARTY, M., et SCHAFER, M. (1975). Estimating public opinion with the randomized response model. *Public Opinion Quarterly*, 39, 507-513.
- ZDEP, S.M., et RHODES, I.N. (1976). Making the randomized response technique work. *Public Opinion Quarterly*, 40, 531-537.
- ZDEP, S.M., RHODES, I.N., SCHWARZ, R.M., et KILKENNY, M.J. (1979). The validity of the randomized response technique. *Public Opinion Quarterly*, 43, 544-549.

- SAKASEGAWA, H., et TAKAHASI, K. (1974). Effects of repetition and finite population corrections in randomized response models (en japonais). *Proceedings of the Institute of Statistical Mathematics*, 22, 59-67.
- SAKASEGAWA, H., TAKAHASI, K., et SUZUKI, T. (1977). An investigation of a new randomized response model (en japonais). *Proceedings of the Institute of Statistical Mathematics*, 25.
- AER SCHEERS, N.J., et DAYTON, C.M. (1982). The covariate unrelated question randomized response model. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 407-410.
- H SCHEERS, N.J., et DAYTON, C.M. (1986). RRCOV: Computer program for covariate randomized response models. *American Statistician*, 40, 229.
- AHV SCHMEIDLER, J. (1987). Assessing quality of randomized response: were instructions followed? *Proceedings of the Survey Research Methods Section, American Statistical Association*, 245-249.
- HQT SEN, P.K. (1974a). On the estimation of basic distributions in randomized response models. *Communications in Statistics*, 3, 1081-1092.
- EHT SEN, P.K. (1974b). On unbiased estimation for randomized response models. *Journal of the American Statistical Association*, 69, 997-1001.
- HRT SEN, P.K. (1976). Asymptotically optimal estimators of general parameters in randomized response models (with discussion). *International Statistical Review*, 44, 223-224.
- AE SHIMIZU, I.M., et BONHAM, G.S. (1978). Randomized response technique in a national survey. *Journal of the American Statistical Association*, 73, 35-39.
- AE SHOTLAND, R.L., et YANKOVSKI, L.D. (1982). The random response method: a valid and ethical indicator of the "truth" in reactive situations. *Personality and Social Psychology Bulletin*, 8, 174-179.
- P SILVA, L.C. (1983). The randomized response technique: a general model for polytomous variables (en espagnol). *Investigation Operational*, 4, 75-100.
- RT SIMMONS, W.R. (1970). Response to randomized inquiries: a technique for reducing bias. *Administrative Applications Division Conference Transactions, American Society for Quality Control*, Chapter 4-B.
- EH SINGH, J. (1976). A note on the randomized response technique. *Proceedings of the Social Statistics Section, American Statistical Association*, 772.
- EH SINGH, J. (1978). A note on maximum likelihood from randomized response models. *Proceedings of the Social Statistics Section, American Statistical Association*, 282-283.
- AX SMITH, E.P., et SOSNOWSKI, T.S. (1972). Faculty evaluations by randomized response sampling. *Journal of Experimental Education*, 41, 70-72.
- AE SMITH, L.L., FEDERER, W.T., et RAGHAVARAO, D. (1974). A comparison of three techniques for eliciting truthful answers to sensitive questions. *Proceedings of the Social Statistics Section, American Statistical Association*, 447-452.
- BX SOEKEN, K.L. (1987). Randomized response methodology in health research. *Evaluation and Health Professions*, 10, 58-66.
- AEH SOEKEN, K.L., et MACREADY, G.B. (1982). Respondents' perceived protection when using randomized response. *Psychological Bulletin*, 92, 487-489.
- ACE SOEKEN, K.L., et MACREADY, G.B. (1985). Randomized response parameters as factors in frequency estimates. *Educational and Psychological Measurement*, 45, 89-100.
- HT SPURRIER, J.D., et PADGETT, W.J. (1980). The application of Bayesian techniques in randomized response. Dans *Sociological Methodology*, San-Francisco: Jossey-Bass, 533-544.

OLIVIERI, D., et BRESSAN, F. (1984). A randomized response model with fixed alternative answer (en italien). *Rivista di Statistica Applicata*, 17, 165-172.

ORWIN, R.G., et BORUCH, R.F. (1982). RRT meets RDD: statistical strategies for assuring response privacy in telephone surveys. *Public Opinion Quarterly*, 46, 560-571.

C O'BRIEN, D.M., et COCHRAN, R.S. (1977). The comprehension factor in randomized response. *Proceedings of the Social Statistics Section, American Statistical Association*, 270-272.

O'BRIEN, D.M., et COCHRAN, R.S. (1978). The effect of less than complete truthfulness on a quantitative randomized response model. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 743-747.

HT O'HAGAN, A. (1987). Bayes linear estimators for randomized response models. *Journal of the American Statistical Association*, 82, 584-587.

AE PEARL, R.L., et FEDERER, W.T. (1975). Varying levels of probability for selecting sensitive questions using a randomized response technique. *Proceedings of the Social Statistics Section, American Statistical Association*, 584-587.

T PITZ, G.E. (1980). Bayes analysis of random response models. *Psychological Bulletin*, 87, 209-212.

X POHL, B.B., et POHL, N.F. (1975). Random response techniques for reducing non-sampling error in interview survey research. *Journal of Experimental Education*, 44, 48-53.

EQR POLLOCK, K.H., et BEK, Y. (1976). A comparison of three randomized response models for quantitative data. *Journal of the American Statistical Association*, 71, 884-886.

QR POOLE, W.K. (1974). Estimation of the distribution function of a continuous type random variable through randomized response. *Journal of the American Statistical Association*, 69, 1002-1005.

MQ POOLE, W.K., et CLAYTON, C.A. (1982). Generalizations of a contamination model for continuous type random variables. *Communications in Statistics, Sér. A*, 11, 1733-1742.

EH RAGHAVARAO, D. (1978). On an estimation problem in Warner's randomized response technique. *Biometrics*, 34, 87-90.

ER RAGHAVARAO, D., et FEDERER, W.T. (1979). Block total response as an alternative to the randomized response method in surveys. *Journal of the Royal Statistical Society, Sér. B*, 41, 40-45.

AE REASER, J.M., HARTSOCK, S., et HOEHN, A.J. (1975). A test of the forced alternative random response questionnaire technique. Technical Report 75-9, Human Resources Research Organization, Alexandria, Virginia.

AEQ REINMUTH, J.E., et GEURTS, M.D. (1975). The collection of sensitive information using a two-stage randomized response model. *Journal of Marketing Research*, 12, 402-407.

CM ROSENBERG, M.J. (1979). Multivariable analysis of a randomized response technique for statistical disclosure control. Thèse de doctorat, University of Michigan, Ann Arbor.

AEHP ROSENBERG, M.J. (1980). Categorical data analysis by a randomized response technique for statistical disclosure control (with discussion). *Proceedings of the Survey Research Methods Section, American Statistical Association*, 311-316.

H ROSENBERG, M.J. (1985). An application of PROC FUNCAT to randomized response data. *Proceedings of the SAS Users Group International Conference*, 10, 1070-1075.

AE ROSENBLATT, R.R., et KELLY, E.L. (1978). A comparison of the sensitivity of the unrelated question randomized response model with three other data accumulation techniques using examination cheating as a model. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 356-361.

- MARASINI, D. (1978). La stratificazione nel campionamento con riposte casualizzate. *Statistica*, 38, 493-506.
- MARASINI, D. (1981). The randomized response in the two-stage sampling scheme (en italien). *Quaderni di Statistica e Matematica Applicata*, 4, 81-96.
- MARASINI, D., et FERRARI, P. (1983). Sampling with randomized responses: estimation and hypotheses testing in case of stratified and two-stage sampling (en italien). *Rivista di Statistica Applicata*, 16, 15-41.
- MARASINI, D., et OLIVIERI, D. (1983). Randomized response models and the quality of statistical data (en italien). *Società Italiana di Statistica, Atti del Convegno*, 1, 489-513.
- MATLOFF, N.S. (1984). Use of covariates in randomized response settings. *Statistics and Probability Letters*, 2, 31-34.
- MAZZALI, A. (1983). A scheme of sampling with randomized responses in case of k questions (en italien). *Rivista di Statistica Applicata*, 16, 99-105.
- MCDANIEL, S.A., et KROTKI, K.J. (1979). Estimates of the rate of illegal abortion and the effect of eliminating therapeutic abortion, Alberta 1973-74. *Canadian Journal of Public Health*, 70, 393-398.
- MILLER, J.D. (1984). A new survey technique for studying deviant behavior. Thèse de doctorat, George Washington University, Washington, DC.
- MOORS, J.J.A. (1971). Optimization of the unrelated question randomized response model. *Journal of the American Statistical Association* 66, 627-629.
- MOORS, J.J.A. (1981). Inadmissibility of linearly invariant estimators in truncated parameter spaces. *Journal of the American Statistical Association*, 76, 910-915.
- MOORS, J.J.A. (1985). Estimation in truncated parameter spaces. Thèse de doctorat, Université de Tilburg, Pays-Bas.
- MORIARTY, M., et WISEMAN, F. (1976). On the choice of a randomization technique with the randomized response model. *Proceedings of the Social Statistics Section, American Statistical Association*, 62-4-626.
- MUKERJEE, R. (1981). Inference on confidential characters from survey data. *Calcutta Statistical Association Bulletin*, 30, 77-88.
- MUKHOPADHYAY, P. (1980a). A survey on the socio-economic conditions of some college students of Calcutta. Rapport non-publié, Indian Statistical Institute.
- MUKHOPADHYAY, P. (1980b). On estimation of some confidential characters from survey data. *Calcutta Statistical Association Bulletin*, 29, 133-141.
- MUKHOPADHYAY, P., and HALDER, A.K. (1980). Bayesian tables for Warner's randomized response probabilities. Technical Report ASC802, Indian Statistical Institute, Calcutta.
- OLIVIERI, D. (1981). Le risposte casualizzate: stime, dimensioni ed errori campionari. *Rivista di Statistica Applicata*, 14, 79-98.
- OLIVIERI, D. (1982). *La diffusione della droga nelle scuole secondarie superiori di Verona*. Vicenza, Italie: Cassa di Risparmio di Verona Vicenza e Belluno.
- OLIVIERI, D. (1983a). On a modification of Simon's scheme of sampling with randomized responses, with efficiency comparisons (en italien). *Rivista di Statistica Applicata*, 16, 57-75.
- OLIVIERI, D. (1983b). Stratified sampling with randomized responses and fixed alternative response (en italien). *Rivista di Statistica Applicata*, 16, 77-84.
- OLIVIERI, D. (1984). Estimation of parameters and efficiency in the Poole's randomized response model (en italien). *Società Italiana di Statistica, Atti della Riunione Scientifica*, 32, 463-472.

- LANDENNA, G. (1983). Sampling with randomized responses: a general view (en italien). *Rivista di Statistica Applicata*, 16, 5-14.
- LANKE, J. (1975). On the choice of the unrelated question in Simon's version of randomized response. *Journal of the American Statistical Association*, 70, 80-83.
- LANKE, J. (1976). On the degree of protection in randomized interviews (with discussion). *International Statistical Review*, 44, 197-203.
- LAVIN, P. (1974). A necessary and sufficient condition for asymptotic masking of the Warner MLE estimate. Report No. 79, Institute of Statistics, Université de Stockholm.
- LEARNER, M. (1973). The collection of data on deviant behavior: public policy issues (résumé). *Bulletin of the International Statistical Institute*, 45, 150.
- LEVY, K.J. (1976a). Reducing the occurrence of omitted or untruthful responses when testing hypotheses concerning proportions. *Psychological Bulletin*, 83, 759-761.
- LEVY, K.J. (1976b). The randomized response technique and large sample pairwise comparisons among the parameters of k independent binomial populations. *British Journal of Mathematical and Statistical Psychology*, 29, 257-262.
- LEVY, K.J. (1977a). The randomized response technique and appropriate sample sizes for selecting the largest value of π from among k binomial populations. *British Journal of Mathematical and Statistical Psychology*, 30, 234-236.
- LEVY, K.J. (1977b). The randomized response technique and comparisons among the parameters of k independent binomial populations. *Psychological Bulletin*, 84, 244-246.
- LEVY, K.J. (1978). Sample size comparisons involving the randomized response technique. *Journal of Experimental Education*, 47, 21-23.
- LEVY, K.J. (1980). The randomized response technique and large sample tests concerning the parameters of a multinomial distribution. *Educational and Psychological Measurement*, 40, 701-708.
- LEYSIEFFER, F.W. (1975). Respondent jeopardy in randomized response procedures. Technical Report M338, Florida State University, Tallahassee, Florida.
- LEYSIEFFER, F.W., et WARNER, S.L. (1976). Respondent jeopardy and optimal designs in randomized response models. *Journal of the American Statistical Association*, 71, 649-656.
- LIU, P.T., et CHOW, L.P. (1976a). A new discrete quantitative randomized response model. *Journal of the American Statistical Association*, 71, 72-73.
- LIU, P.T., et CHOW, L.P. (1976b). The efficiency of the multiple trial randomized response technique. *Biometrics*, 32, 607-618.
- LIU, P.T., CHEN, C.N., and CHOW, L.P. (1976). A study of the feasibility of Hopkins randomized response models. *Proceedings of the Social Statistics Section, American Statistical Association*, 561-566.
- LIU, P.T., CHOW, L.P., et MOSTLEY, H.W. (1975). Use of the randomized response technique with a new randomizing device. *Journal of the American Statistical Association*, 70, 329-332.
- LOCANDER, W., SUDMAN, S., et BRADBURN, N. (1976). An investigation of interview method, threat and response distortion. *Journal of the American Statistical Association*, 71, 269-275.
- LOYNES, R.M. (1976). Asymptotically optimal randomized response procedures. *Journal of the American Statistical Association*, 71, 924-928.
- MADIGAN, F.C., ABERNATHY, J.R., HERRIN, A.N., et TAN, C. (1976). Purposive concealment of death in household surveys in Misamis Oriental Province. *Population Studies*, 30, 295-303.

- HORVITZ, D.G., GREENBERG, B.G., et ABERNATHY, J.R. (1976a). Randomized response: a data gathering device for sensitive questions (with discussion). *International Statistical Review*, 44, 181-196.
- HORVITZ, D.G., GREENBERG, B.G., et ABERNATHY, J.R. (1976b). The randomized response technique. Dans *Perspectives on Attitude Assessment: Surveys and their Alternatives*, (eds. H.W. Sinaiko et L.A. Broedling), Champaign: Pendelton Publications.
- AER HORVITZ, D.G., SHAH, B.V., et SIMMONS, W.R. (1967). The unrelated question randomized response model. *Proceedings of the Social Statistics Section, American Statistical Association*, 65-72.
- IGLEWITZ, B. (1976). A coding approach to the sensitive question problem. *Proceedings of the Social Statistics Section, American Statistical Association*, 414-415.
- ER IIT Research Institute and the Chicago Crime Commission (1971). A study of organized crime in Chicago. IITRI Project No. H-6031, rapport préparé pour l'Illinois Enforcement Commission, Chicago.
- A KAMMERMANN, L.A., GREENBERG, B.G., et QUADE, D. (1985). Selecting optimal values for π_j in the unrelated question randomized response model. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 470-475.
- CO KIM, J.J. (1978). Randomized response techniques for surveying human populations. Thèse de doctorat, Temple University, Philadelphie.
- MP KIM, J.J. (1987). A further development of randomized response for masking dichotomous variables. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 239-244.
- MR KIM, J.J., et FLUECK, J.A. (1976). A review of randomized response designs and some new results. *Proceedings of the Social Statistics Section, American Statistical Association*, 477-482.
- BM KIM, J.J., et FLUECK, J.A. (1978a). An additive randomized response model. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 351-355.
- AEMP KIM, J.J., et FLUECK, J.A. (1978b). Modifications of the randomized response technique for sampling without replacement. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 346-350.
- AEH KOCH, G.G., ABERNATHY, J.R., et IMREY, P.B. (1975). On a method for study of family size preferences. *Demography*, 12, 57-66.
- X KOLATA, G. (1987). How to ask about sex and get honest answers. *Science*, 236, 382.
- MP KRAEMER, H.C. (1980). Estimation and testing of bivariate association using data generated by the randomized response technique. *Psychological Bulletin*, 87, 304-308.
- AEV KROTKE, K.J., et FOX, B. (1974). The randomized response technique, the interview and the self administered questionnaire: an empirical comparison of fertility reports. *Proceedings of the Social Statistics Section, American Statistical Association*, 367-371.
- AE KROTKE, K.J., et McDANIEL, S.A. (1975). Three estimates of illegal abortion in Alberta, Canada: Survey, mailback questionnaire and randomized response technique. *Bulletin of the International Statistical Institute*, 46, 67-70.
- A KROTKE, K.J., et McDANIEL, S.A. (1978). La technique de réponse rendue aléatoire; quelques résultats d'une étude à Edmonton, Alberta. *Population et Famille*, 41, 91-119.
- B LAI, C.D. (1982). *A review of randomized response survey models*. Occasional Publications in Mathematics, 12, Department of Mathematics, Massey University, Nouvelle-Zélande.
- AEQV LAMB, C.W., et STEM, D.E. (1978). An empirical validation of the randomized response technique. *Journal of Marketing Research*, 15, 616-621.

GOODSTADT, M.S., et GRUSON, V. (1975). The randomized response technique: a test on drug use. *Journal of the American Statistical Association*, 70, 814-818.

AE
use: the randomized response technique. *International Journal of Addictions*, 13, 359-367.

GOULD, A.L., SHAH, B.V., et ABERNATHY, J.R. (1969). Unrelated question randomized response techniques with two trials per respondent. *Proceedings of the Social Statistics Section, American Statistical Association*, 351-359.

AQ
randomized response technique in obtaining quantitative data. *Proceedings of the Social Statistics Section, American Statistical Association*, 40-43.

X
technique and application to the field of public health. *Millbank Memorial Fund Quarterly*, 484, Pt. 2, 39-55.

BX
response. Dans *Encyclopedia of Statistical Sciences* (Vol. 7), (eds. S. Kotz, N.L. Johnson et C.B. Read), New York: John Wiley, 540-548.

ORST
The unrelated question randomized response model: theoretical framework. *Journal of the American Statistical Association*, 64, 520-539.

EHR
randomized response designs. Dans *Reliability and Biometry: Statistical Analysis of Lifelength*, (eds. F. Proschan et R.J. Serfling), Philadelphia: SIAM, 787-815.

EHQR
Application of the randomized response technique in obtaining quantitative data. *Journal of the American Statistical Association*, 66, 243-250.

CE
Respondent hazards in the unrelated question randomized response model. *Journal of Statistical Planning and Inference*, 1, 53-60.

EHT
models. *Communications in Statistics*, Sér. A, 14, 2411-2435.

EHT
GUNEL, E. (1985b). On the design of randomized response sampling plan. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 457-459.

CT
HAYASHI, C. (1968). Response errors and biased information. *Annals of the Institute of Statistical Mathematics*, 20, 211-228.

C
HILMAR, N.A. (1968). Anonymity, confidentiality and invasions of privacy: the responsibility of the researcher. *American Journal of Public Health*, 58, 324-330.

EQR
response technique for eliminating evasiveness to quantitative response questions. *Psychological Bulletin*, 87, 525-530.

BMQ
HIMMELFARB, S., et EDGELL, S.E. (1982). Note on "The randomized response approach." Addendum to Fox and Tracy. *Evaluation Review*, 6, 279-284.

C
HIMMELFARB, S., et LICKTEIG, C. (1982). Social desirability and the randomized response technique. *Journal of Personality and Social Psychology*, 43, 710-717.

EHPR
HOCHBERG, Y. (1975). Two stage randomized response schemes for estimating a multinomial. *Communications in Statistics*, 4, 1021-1032.

B
HORVITZ, D.G., GREENBERG, B.G., et ABERNATHY, J.R. (1975). Recent developments in randomized response designs. Dans *A Survey of Statistical Design and Linear Models*, (ed. J.N. Srivastava), New-York: North Holland, 271-285.

- ERIKSSON, S.A. (1976a). Some sampling theory for surveys with randomized response interviews. Report No. 8, Confidentiality in Surveys Research Project, Institute of Statistics, Université de Stockholm.
- ERIKSSON, S.A. (1976b). Regression analysis of data from randomized interviews. Report No. 17, Confidentiality in Surveys Research Project, Institute of Statistics, Université de Stockholm.
- H ERIKSSON, S.A. (1978). Il test sequenziale di Wald nel campionamento con risposte casualizzate. *Statistica*, 38, 481-492.
- S FERRARI, P. (1984). Two-stage sampling with randomized response and unknown strata sizes (en italien). *Quaderni di Statistica e Matematica Applicata*, 5, 5-19.
- AE FIDLER, D.S., et KLEINKNECHT, R.E. (1977). Randomized response versus direct questioning: two data collection methods for sensitive information. *Psychological Bulletin*, 84, 1045-1049.
- MX FIERING, M.B., et HOOVER, M. (1985). Analysis of disclosure avoidance procedures. *Civil Engineering Systems*, 2, 12-19.
- CEHR FLIGNER, M.A., POLICELLO, G.E., et SINGH, J. (1977). A comparison of two randomized response survey methods with consideration for the level of respondent protection. *Communications in Statistics*, Sér. A, 6, 1511-1524.
- B FLUECK, J.A., et KIM, J.J. (1976). *Bibliography for randomized response*. Mimeo Series No. 33, Department of Statistics, Temple University.
- AV FOLSOM, R.E. (1974). A randomized response validation study: comparison of direct and randomized reporting of DUI arrests. Final Report 25U-807, Research Triangle Institute, Research Triangle Park, Caroline du Nord.
- AEHR FOLSOM, R.E., GREENBERG, B.G., HORVITZ, D.G., et ABERNATHY, J.R. (1973). The two alternate questions randomized response model for human surveys. *Journal of the American Statistical Association*, 68, 525-530.
- AHOV FOX, J.A., et TRACY, P.E. (1980a). A field-validation of a quantitative randomized response model (with discussion). *Proceedings of the Survey Research Methods Section, American Statistical Association*, 299-304.
- A FOX, J.A., et TRACY, P.E. (1980b). The randomized response approach: applicability to criminal justice research and evaluation. *Evaluation Review*, 4, 601-622.
- H FOX, J.A., et TRACY, P.E. (1984). Measuring associations with randomized response. *Social Science Research*, 13, 188-197.
- BX FOX, J.A., et TRACY, P.E. (1986). *Randomized response: a method for sensitive surveys*. Beverly Hills: Sage Publications.
- AV GERSTEL, E.K., BRUCE, J., FOLSOM, R.E., et DURHAM, J. (1974). The effectiveness of the Mecklenburg county alcohol safety action project. Rapport non-publié, Research Triangle Institute, Research Triangle Park, Caroline du Nord.
- AV GERSTEL, E.K., MOORE, P., FOLSOM, R.E., et KING, D.A. (1970). Mecklenburg county drinking driving attitude survey. Rapport non-publié, Research Triangle Institute, Research Triangle Park, Caroline du Nord.
- A GEURTS, M.D., ANDRUS, R.R., et REINMUTH, J.E. (1975). Researching shopping and other deviant customer behavior using the randomized response design. *Journal of Retailing*, 51, 43-48.
- HOT GODAMBE, V.P. (1980). Estimation in randomized response trials. *International Statistical Review*, 48, 29-32.
- A GOODE, T., et HEINE, W. (1978). Surveying the extent of drug use. *Survey Statistician*, 1, 10-12.

- DAWES, R.M., and SMITH, T.L. (1985). Attitude and opinion measurement. Dans *Handbook of Social Psychology*, (3^e éd.), (éds. G. Lindzey, et E. Aronson), Hillsdale: Erlbaum, 509-566.
- DEFAA, W. (1982). *Anonymisierte Befragungen mit zufallsverschlüsselten Antworten: Die Randomized-Response-Technik (RRT)*. Frankfurt am Main: Verlag Peter Lang.
- HT DELACÉY, P.W. (1975). Randomized conditional response. *Proceedings of the Social Statistics Section, American Statistical Association*, 383-386.
- T DEVORE, J.L. (1977). A note on the randomized response technique. *Communications in Statistics*, Sér. A, 6, 1525-1534.
- X DEVORE, J.L. (1979). Estimating a population proportion using randomized responses. *Mathematics Magazine*, 52, 38-40.
- EHO DOWLING, T.A., et SHACHTMAN, R.H. (1975). On the relative efficiency of randomized response models. *Journal of the American Statistical Association*, 70, 84-87.
- H DOWNS, T., GILLAND, D.C., et KATZ, L. (1978). Probability in a contested election. *American Statistician*, 32, 122-125.
- EHQ DRAGO, E. (1981). Estimate of the mean and the second moment of a population through randomized response sampling (en italien). *Rivista di Matematica per le Scienze Economiche e Sociali*, 4, 49-58.
- HM DRANE, W. (1975). Randomized response to more than one question. *Proceedings of the Social Statistics Section, American Statistical Association*, 395-397.
- HM DRANE, W. (1976). On the theory of randomized responses to two sensitive questions. *Communications in Statistics*, Sér. A, 5, 565-574; Corrigenda (1976), 5, 1552.
- HQ DUFFY, J.C., et WATERTON, J.J. (1984). Randomized response models for estimating the distribution function of a quantitative character. *International Statistical Review*, 52, 165-172.
- E DURHAM, A.M., et LICHTENSTEIN, M.J. (1983). Response bias in self-report surveys: evaluation of randomized responses. Dans *Measurement Issues in Criminal Justice*, (éd. G.P. Waldo), Beverly Hills: Sage Publications.
- Q EDGELL, E. (1980). Additive constants model: a randomized response technique for eliminating evasiveness to quantitative response. *Psychological Bulletin*, 87, 304-308.
- EHO EDGELL, S.E., HIMMELFARB, S., et CIRA, D.J. (1986). Statistical efficiency of using two quantitative randomized response techniques to estimate correlation. *Psychological Bulletin*, 100, 251-256.
- AV EDGELL, S.E., HIMMELFARB, S., et DUCHAN, K.L. (1982). Validity of forced responses in a randomized response model. *Sociological Methods and Research*, 11, 89-100.
- EHQ EICHHORN, B., et HAYRE, L.S. (1983). Scrambled randomized response methods for obtaining sensitive quantitative data. *Journal of Statistical Planning and Inference*, 7, 307-316.
- X ELEM, D., et HOWSON, A. (1979). Randomized response: a survey technique for sensitive questions. Dans *Interactive Statistics*, (éd. D. McNiel), New-York: Elsevier North-Holland, 193-207.
- B EMRICH, L. (1983). Randomized response techniques. Dans *Incomplete Data in Sample Surveys*, Vol. 2, (éds. W.G. Madow, I. Olkin et D.B. Rubin), New York: Academic Press, 73-80.
- EHQ ERIKSSON, S.A. (1973a). A new model for randomized response. *International Statistical Review*, 41, 101-113.
- QR ERIKSSON, S.A. (1973b). Randomized interviews for sensitive questions. Thèse de doctorat, Université de Göteborg, Suède.

- CHAUDHURI, A., et ADHIKARI, A.K. (1981). Sampling strategies with randomized response trials and their properties and relative efficiencies. Technical Report ASC815, Indian Statistical Institute, Calcutta.
- CHAUDHURI, A., et MUKERJEE, R. (1985). Optionally randomized response techniques. *Calcutta Statistical Association Bulletin*, 34, 225-229.
- CHAUDHURI, A., et MUKERJEE, R. (1987). Randomized response techniques: a review. *Statistica Neerlandica*, 41, 27-44.
- CHAUDHURI, A., et MUKERJEE, R. (1988). *Randomized response: theory and techniques*. New York: Marcel Dekker.
- CHEN, E., CHOW, L.P., et LIU, P.T. (1974). Field studies on the new randomized response techniques. Department of Population Dynamics, Johns Hopkins University, Baltimore.
- CHEN, T.T. (1978). Log-linear models for the categorical data obtained from randomized response techniques. *Proceedings of the Social Statistics Section, American Statistical Association*, 284-288.
- CHEN, T.T. (1979). Analysis of randomized response as purposively misclassified data. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 158-163.
- CHI, I.C., CHOW, L.P., et RIDER, R.V. (1972). The randomized response techniques as used in the Taiwan outcome and pregnancy study. *Studies in Family Planning*, 3, 265-269.
- CHOW, L.P., et LIU, P.T. (1973). A new randomized response technique: the multiple answer model. Department of Population Dynamics, Johns Hopkins University, Baltimore.
- CHOW, L.P., GRUHN, W., et CHANG, W.P. (1979). Feasibility of the randomized response technique in rural Ethiopia. *American Journal of Public Health*, 69, 273-276.
- CLICKNER, R.P., et IGLEWICZ, B. (1980). Warner's randomized response technique: the two sensitive questions case. *South African Statistical Journal*, 14, 77-86.
- COHEN, J.E. (1987). Sexual behavior and randomized responses. *Science*, 236, 1503.
- COMSTOCK, G.W., CONDE, J.G., et HELSING, K.J. (1985). A simple randomized response device. *American Journal of Epidemiology*, 122, 187-190.
- CURLETTE, W.C. (1980). The randomized response technique: using probability theory to ask sensitive questions. *Mathematics Teacher*, 73, 618-621.
- DALENIUS, T. (1977). Privacy transformations for statistical information systems. *Journal of Statistical Planning and Inference*, 1, 73-86.
- DALENIUS, T. (1983). Randomized response. Dans *Solutions to Ethical and Legal Problems in Social Research*, New York: Academic Press, 237-248.
- DALENIUS, T., et VITALE, R.A. (1979). A new randomized response technique for estimating the mean of a distribution. Dans *Contributions to Statistics, Jaroslav Hajek Memorial Volume*, (éd. J. Jurechkova), Dordrecht: D. Reidel, 43-47.
- DANERMARK, B., et SWENSSON, B. (1987). Measuring drug use among Swedish adolescents. *Journal of Official Statistics*, 3, 439-448.
- DANIEL, W.W. (1979). *Collecting sensitive data by randomized response: unannotated bibliography*. Research Monograph No. 85, College of Business Administration, Georgia State University.
- DAWES, R.M., et MOORE, M. (1980). Guttman scaling orthodox and randomized responses (en allemand). Dans *Attitude Measurement*, (éd. F. F. Peterman), Göttingen: Verlag für Psychologie, 117-133.

CE

B

ACE

QR

CX

C

X

RX

BX

HMS

ARV

PR

AERV

H

H

A

BX

X

OR

HT

- BOURKE, P.D. (1982). Randomized response multivariate designs for categorical data. *Communications in Statistics*, Sér. A, 11, 2889-2901.
- BOURKE, P.D. (1983). Randomized response designs with attribute-based randomization. *Statistical Review*, 5, 125-132.
- BOURKE, P.D. (1984). Estimation of proportions using symmetric randomized response designs. *Psychological Bulletin*, 96, 166-172.
- P BOURKE, P.D., et DALENIS, T. (1973). Multi-proportions randomized response using a single sample. Report No. 68, Errors in Surveys Research Project, Institute of Statistics, Université de Stockholm.
- HT BOURKE, P.D., et DALENIS, T. (1974a). A note on inadmissible estimates in randomized enquiries. Report No. 72, Errors in Surveys Research Project, Institute of Statistics, Université de Stockholm.
- HOT BOURKE, P.D., et DALENIS, T. (1974b). Randomized response models with lying. Report No. 71, Errors in Surveys Research Project, Institute of Statistics, Université de Stockholm.
- CMR BOURKE, P.D., et DALENIS, T. (1976). Some new ideas in the realm of randomized inquiries (with discussion). *International Statistical Review*, 44, 219-221.
- HQ BOURKE, P.D., et MORAN, M.A. (1984). Application of the EM algorithm to randomized response data. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 788-793.
- HP BOURKE, P.D., et MORAN, M.A. (1986). An alternative EM formulation for randomized response data. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 444-447.
- ACE BRADBURN, N., et SUDMAN, S. (1979). Improving interview method et questionnaire design. San-Francisco: Jossey-Bass, 1-13.
- B BRESSAN, F. (1983). Warner's scheme of sampling with randomized responses with memory (en italien). *Rivista di Statistica Applicata*, 16, 85-98.
- ACEMV BREWER, K.R.W. (1981). Estimating maritana usage using randomized response - some paradoxical findings. *Australian Journal of Statistics*, 23, 139-148.
- AE BROWN, G.H. (1975). Randomized inquiry vs. conventional questionnaire method in estimating drug usage rates through mail surveys. Technical Report 75-14, Human Resources Research Organization, Alexandria, Virginie.
- AE BROWN, G.H., et HARDING, F.D. (1973). A comparison of methods of studying illicit drug usage. Technical Report 73-9, Human Resources Research Organization, Alexandria, Virginie.
- AEV BUCHMAN, T.A., et TRACY, J.A. (1982). Obtaining responses to sensitive questions: conventional questionnaire versus randomized response technique. *Journal of Accounting Research*, 20, 263-271.
- X CAMPBELL, A.A. (1987). Randomized response technique. *Science*, 236, 1049.
- X CAMPBELL, C., et JOINER, B.L. (1973). How to get the answer without being sure you've asked the question. *American Statistician*, 27, 229-231.
- HO CARR, J.W., MARASCULO, L.A., et BUSK, P. (1982). Optimal randomized response models and methods for hypothesis testing. *Journal of Educational Statistics*, 7, 295-310.
- AH CHAUDHURI, A. (1983). Randomized response technique to determine input in crop estimation. *Calcutta Statistical Association Bulletin*, 32, 208-210.
- OQS CHAUDHURI, A. (1987). Randomized response surveys of finite populations: a unified approach with quantitative data. *Journal of Statistical Planning and Inference*, 15, 157-165.

- BARKSDALE, W.B. (1975). New randomized response techniques for control of non-sampling errors in surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 302-304.
- BARTH, J.T., et SANDLER, H.M. (1976). Evaluation of the randomized response technique in a drinking survey. *Journal of Studies in Alcoholism*, 37, 690-693.
- BASULTO, J. (1982). The randomized response design of Warner: a superpopulation model (en espagnol). *Estadística Española*, 96, 51-61.
- BEGIN, G., BOIVIN, M., et BELLEROSE, J. (1979). Sensitive data collection through the random response technique: some improvements. *Journal of Psychology*, 101, 53-65.
- BELDT, S.F., DANIEL, W.W., et GARCHA, B.S. (1982). The Takahasi-Sakasegawa randomized response technique. A field test. *Sociological Methods and Research*, 11, 101-111.
- BELTHOUSE, D.R. (1980). Linear models for randomized response designs. *Journal of the American Statistical Association*, 75, 1001-1004.
- BERMAN, J., MCCOMBS, H., et BORUCH, R.F. (1977). Notes on the contamination method. *Sociological Methods and Research*, 6, 45-62.
- BLANGIARDO, G.C. (1978). I campioni in blocco con risposta casualizzata. *Rivista di Statistica Applicata*, 11, 89-96.
- BLANGIARDO, G.C. (1979). La stratificazione nei campioni in blocco con risposta casualizzata. *Rivista di Statistica Applicata*, 12, 26-36.
- BLOMQUIST, N., et ERIKSSON, S.A. (1974). A general theory of randomized interviews. Research Report 1974:4, Département de statistique, Université de Göteborg.
- BORUCH, R.F. (1971a). Assuring confidentiality of responses in social research: a note on strategies. *American Sociologist*, 6, 308-311.
- BORUCH, R.F. (1971b). Maintaining confidentiality of data in educational research: a systematic analysis. *American Psychologist*, 26, 413-430.
- BORUCH, R.F. (1972). Relations among statistical methods for assuring confidentiality of social research data. *Social Science Research*, 1, 403-414.
- BORUCH, R.F. (1982). Methods for resolving privacy problems in social research. Dans *Ethical Issues in Social Science Research*, (eds. R.R. Faden, R.J. Wallace, et L. Walters), Baltimore: Johns Hopkins University Press.
- BORUCH, R.F., et CECIL, J.S. (1979). *Assuring the confidentiality of social research data*. Philadelphie: University of Pennsylvania Press.
- BOURKE, P.D. (1974a). Multi-proportions randomized response using the unrelated question technique. Report No. 74, Errors in Surveys Research Project, Institute of Statistics, Université de Stockholm.
- BOURKE, P.D. (1974b). Symmetry of response in randomized response designs. Report No. 75, Errors in Surveys Research Project, Institute of Statistics, Université de Stockholm.
- BOURKE, P.D. (1974c). Vector response in randomized response designs. Report No. 76, Errors in Surveys Research Project, Institute of Statistics, Université de Stockholm.
- BOURKE, P.D. (1975). Randomized response designs for multivariate estimation. Report No. 6, Confidentiality in Surveys Research Project, Institute of Statistics, Université de Stockholm.
- BOURKE, P.D. (1978). Randomized response designs with symmetric response for multi-proportions estimation. *Statistical Review*, 16, 197-204.
- BOURKE, P.D. (1981). On the analysis of some multivariate randomized response designs for categorical data. *Journal of Statistical Planning et Inference*, 5, 165-170.

EMO
AEV
HOQT
CR
S
S
MT
C
C
CE
CX
CX
RT
T
HM
MR
EMR

REMERCIEMENTS

L'auteur souhaite exprimer sa reconnaissance aux nombreux collègues et auteurs d'articles sur la méthode des réponses randomisées qui ont bien voulu relire une version antérieure de cette bibliographie, faire des commentaires et communiquer des références supplémentaires et des réimpressions de leurs articles. L'établissement de cette bibliographie a été rendu possible en partie grâce à une subvention de la National Science Foundation (n° SES-8612320).

BIBLIOGRAPHIE

ABERNATHY, J.R., GREENBERG, B.G., et HORVITZ, D.G. (1970). Estimates of induced abortion in urban North Carolina. *Demography*, 7, 19-29.

ABUL-ELA, A.A. (1966). Randomized response models for sample surveys on human populations. Thèse de doctorat, University of North Carolina, Chapel Hill.

ABUL-ELA, A.A., et ABDEL-HAMIED, S.M. (1984). Randomized response ratio estimates: bias and efficiency. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 794-799.

ABUL-ELA, A.A., et ABDEL-HAMIED, S.M. (1985). A randomized response ratio estimate from quantitative data. *Proceedings of the Social Statistics Section, American Statistical Association*, 300-305.

ABUL-ELA, A.A., et DAKKOURI, H.M. (1980). Randomized response models: a ratio estimator. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 205-208.

ABUL-ELA, A.A., GREENBERG, B.G., et HORVITZ, D.G. (1967). A multi-proportions randomized response model. *Journal of the American Statistical Association*, 62, 990-1008.

ADHIKARI, A.K. (1982). On randomized response surveys with sensitive quantitative characters: a case study in the Indian Statistical Institute. Technical Report ASC826, Indian Statistical Institute, Calcutta.

ADHIKARI, A.K., CHAUDHURI, A., et VIJAYAN, K. (1984). Optimum sampling strategies for randomized response trials (with discussion). *International Statistical Review*, 52, 115-125.

AHSANULLAH, M., et EICHHORN, B.H. (1984). On scrambled response of sensitive quantitative data. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 800-802.

ALBERS, W. (1982). Simple randomized response procedures with bounded respondent risk for quantitative data. *Kwantitative Methoden*, 8, 35-46.

ALEXANDER, J.R. (1978). Probability as an aid in social research: the randomized response technique. *Mathematical Spectrum*, 11, 10-13.

ANDERSON, H. (1975). Efficiency versus protection in randomized response designs. Thèse de doctorat, Université de Lund, Suède.

ANDERSON, H. (1976). Estimation of a proportion through randomized response (with discussion). *International Statistical Review*, 44, 213-217.

ANDERSON, H. (1977). Efficiency versus protection in a general randomized response model. *Scandinavian Journal of Statistics*, 4, 11-19.

BARCKSDALE, W.B. (1971). New randomized response techniques for control of non-sampling errors in surveys. Thèse de doctorat, University of North Carolina, Chapel Hill.

commentée. Trois ouvrages généraux ont également été publiés sur la méthode des réponses randomisées: Defaa (1982), Fox et Tracy (1986) et Chaudhuri et Mukerjee (1988). Malheureusement, aucun ne comporte de bibliographie exhaustive et à jour, et celle que nous proposons ici a pour but de combler cette lacune.

Pour être aussi complet que possible, nous énumérons les textes publiés et non publiés, mais nous avons évidemment pu incorporer ces derniers seulement quand il y avait des renseignements à leur sujet dans les diverses sources. En outre, pour éviter de mentionner deux fois un même document, nous avons exclu les textes non publiés présentés à des conférences et dont le contenu est presque intégralement repris dans une publication subséquente. Les thèses de doctorat sont en général incluses, toutefois, parce qu'elles contiennent habituellement plus de détails que les articles qu'on a pu en tirer. En ce qui concerne les articles traitant de l'utilisation d'autres méthodes que celles des questions randomisées pour obtenir des réponses à des questions délicates, ils figurent dans la bibliographie seulement si l'on y compare ces autres méthodes à celle des réponses randomisées. Les articles sur l'utilisation des techniques de randomisation pour assurer la confidentialité de données déjà recueillies (par exemple l'arrondissement ou le codage aléatoire) ne sont pas inclus à moins qu'il n'y soit également question de randomisation dans le processus de collecte proprement dit.

Notre bibliographie est un répertoire alphabétique qui donne pour chaque entrée tous les renseignements bibliographiques habituels. Le titre de l'article ou de l'ouvrage dans la langue d'origine est indiqué s'il est connu. Autrement, pour les publications dans des langues autres que l'anglais, nous donnons le titre anglais en indiquant la langue d'origine entre parenthèses. La plupart des textes qui ne sont pas écrits en anglais sont accompagnés d'un résumé en anglais. Une lettre-code correspondant à un classement thématique se trouve dans la marge droite de la page à côté de chaque référence. On trouvera ci-après la liste de ces codes et des catégories qui leur correspondent. On pourra, en s'adressant à l'auteur, se procurer une liste thématique des entrées et un index des auteurs, que le manque d'espace nous empêche de reproduire ici.

2. CODES DES CATEGORIES DE SUJETS

- A - Applications et expériences sur le terrain.
- B - Bibliographies et articles sur des enquêtes.
- C - Confidentialité, attitude des répondants devant la méthode, la manière de comprendre la méthode et la protection de l'anonymat.
- E - Evaluation de techniques ou d'estimateurs de remplacement.
- H - Tests d'hypothèse, estimation et analyse.
- M - Analyse multidimensionnelle.
- O - Optimisation des paramètres du plan.
- P - Questions à choix multiple.
- Q - Variables quantitatives.
- R - Mécanismes et techniques de randomisation.
- S - Plan de sondage.
- T - Développement théoriques.
- V - Etudes de validation.
- X - Exposés descriptifs.

Bibliographie de la méthode des réponses randomisées: 1965-1987

GAD NATHAN¹

RÉSUMÉ

Bibliographie complète des ouvrages, comptes rendus de recherche et articles publiés sur la théorie, l'application et le développement des techniques associées à la méthode des réponses randomisées, avec une classification par sujet.

MOTS CLÉS: Enquête; questions délicates; confidentialité.

1. INTRODUCTION

L'augmentation récente du besoin de données très détaillées sur des sujets délicats (par exemple sur le comportement sexuel, en vue de l'étude de la propagation du sida) a provoqué un réexamen des méthodes pouvant servir à obtenir des réponses à des questions gênantes. On connaît bien les difficultés que soulève l'application des techniques d'enquête classiques à la collecte de renseignements de ce genre, et plusieurs solutions de rechange ont été proposées (Bradburn et Sudman 1979). La principale est la méthode des réponses randomisées, proposée pour la première fois par Warner en 1965. Le principe de cette méthode est que l'enquêteur choisit au moyen d'un mécanisme aléatoire la question à laquelle il va répondre et que l'intervieweur connaît seulement cette réponse et ignore à quelle question elle correspond. Cette technique est censée réduire les biais créés par la non-réponse et l'erreur de réponse du fait qu'elle fait comprendre à l'enquête que l'anonymat de sa réponse est protégé (puisque l'intervieweur ignore la question) et l'incite par conséquent à se montrer plus coopératif et à répondre plus franchement que si une question directe lui était posée.

Depuis 1965, il s'est fait beaucoup de recherche sur les différents aspects de cette méthode: développements théoriques, élaboration de nouvelles techniques de randomisation, applications aux variables quantitatives, aux questions à choix multiple et à l'analyse multidimensionnelle. On a également étudié les problèmes qui touchent l'estimation, l'optimisation des paramètres et le plan de sondage. De nombreuses études empiriques faisant appel à la méthode des réponses randomisées ont été faites sur plusieurs sujets comme la consommation de drogues, l'avortement, l'ivresse au volant et la criminalité; beaucoup de ces études contenaient une évaluation de leurs résultats, faite, dans plusieurs cas, au moyen d'une analyse de validité. Les résultats de ces études étaient très divers, certains révélant des gains sensibles attribuables à la méthode des réponses randomisées, d'autres n'ayant démontré aucune amélioration du taux de réponse ou de la fiabilité des réponses. Pour tâcher d'expliquer les différences dans les résultats empiriques, on a aussi étudié l'attitude des répondants à l'égard des réponses randomisées, leur manière de comprendre cette méthode et l'idée qu'ils se font de la confidentialité et de la protection qu'en permet la méthode.

L'ensemble de ces travaux de recherche est formé de plus de deux cent cinquante thèses, comptes rendus, articles et livres écrits (dans au moins sept langues) depuis une vingtaine d'années. Ce corpus comprend de nombreux exposés descriptifs et études d'enquêtes ainsi que deux bibliographies, celle de Kim et Flueck (1976) et celle de Daniel (1979), cette dernière

¹ Gad Nathan, Département de statistique, Hebrew University, Mt Scopus 91905 Jérusalem, Israël. L'auteur a établi cette bibliographie pendant son séjour au U.S. National Center for Health Statistics, 3700 East-West Highway, Hyattsville, MD 20782.

Comme il en a déjà été fait mention, l'utilisateur qui formule une instruction SQL n'est pas tenu de décrire la façon dont les données sont stockées dans la base de données ou l'emplacement qu'elles y occupent. De ce fait, dans le contexte d'une base de données réparties où SQL est utilisé comme langage d'interrogation de cette base de données, on peut transférer les données entre les machines sans pour autant que les applications actuelles subissent quelque changement que ce soit. SQL devient donc très populaire auprès des réalisateurs de systèmes de gestion de bases de données réparties.

Pour des raisons similaires, SQL gagne en popularité comme langage d'interrogation des machines bases de données. Celles-ci bénéficient du caractère homogène inhérent aux structures de données relationnelles (c'est-à-dire tabulaires) pour les subdiviser sur plusieurs processeurs parallèles. Ces processeurs comportent des jeux d'instructions précisément conçus pour exécuter des opérations relationnelles. En raison du manque de détails quant à la représentation dans les interrogations formulées en SQL, les utilisateurs ne peuvent prendre conscience du rôle joué à l'arrière-plan par ces machines.

4. RÉSUMÉ

Il n'y a pas lieu de s'étonner que SQL soit vite devenu le principal langage d'interrogation de bases de données. Avant longtemps, la très grande majorité des systèmes de gestion de bases de données comporteront une interface faisant appel à SQL. Il est toutefois intéressant de signaler qu'un phénomène insolite se produira; en effet, au fil des années, l'utilisateur «verra» de moins en moins de SQL. Plutôt que de tenter de faire de SQL lui-même un langage à la portée des utilisateurs, on s'efforcera d'élaborer des outils propres à certaines applications qui mettront à la disposition de l'utilisateur une interface expressément conçue en fonction de la tâche à exécuter. SQL deviendra ainsi l'interface la plus courante entre ces outils et les diverses bases de données.

BIBLIOGRAPHIE

- DATE, C.J. (1985). *An Introduction to Database Systems*, Don Mills: Addison-Wesley.
- ORGANISATION INTERNATIONALE DE NORMALISATION (ISO) (1987). *Database Language SQL*, International Standards Organization 9075 (E).


```
SELECT AVG(NO_DE_PERSONNES), MAX(NO_DE_PERSONNES),
MIN(NO_DE_PERSONNES), SUM(NO_DE_PERSONNES),
COUNT(NO_DE_PERSONNES)
FROM LOGEMENT
GROUP BY ENDROIT;
```

Certains fournisseurs y ont ajouté d'autres fonctions telles que la détermination de la variance (VARIANCE) et de l'écart-type (STDDEV). Grâce à ces fonctions supplémentaires, il devient facile de repérer les valeurs aberrantes, dont l'écart par rapport à la moyenne est supérieur à l'écart-type :

```
SELECT LOGEMENT_ID, FROM LOGEMENT
WHERE NO_DE_PERSONNES <
(SELECT AVG(NO_DE_PERSONNES - STDDEV (NO_DE_PERSONNES)
OR
NO_DE_PERSONNES >
(SELECT AVG(NO_DE_PERSONNES) + STDDEV(NO_DE_PERSONNES)
FROM LOGEMENT);
```

3.4 Interface simple avec la base de données

Lorsqu'on interroge une base de données au moyen d'un programme écrit dans un langage hôte comme PL/I, FORTRAN ou C, on se sert aussi d'instructions SQL. Ces instructions sont à toutes fins utiles identiques à celles qu'on emploie quand on interroge en interaction une base de données par le truchement d'un compilateur SQL. La seule différence réside dans le fait que l'interface faisant appel au langage hôte exige une instruction «INTO» supplémentaire pour indiquer les variables du programme qui contiendront les résultats de l'interrogation. En se servant d'une interface identique avec un langage de programmation hôte, on peut diviser en deux activités distinctes la réalisation et la mise au point du programme:

- la mise à l'essai des instructions visant l'extraction/le stockage de données dans la base de données (c'est-à-dire les instructions SQL elles-mêmes), et
- la mise à l'essai du code du programme qui manipule les données.

La première de ces activités peut être accomplie au moyen d'un interpréteur de commandes SQL avant même la rédaction du programme en langage hôte. Les instructions SQL optimales peuvent alors être transférées directement dans le programme hôte, tandis que la mise à l'essai peut être concentrée sur la logique liée à la manipulation des données. Puisque les instructions SQL intégrées au langage hôte sont interprétées au moment de l'exécution, tout changement apporté à l'organisation ou à la structure de la base de données se trouve aussitôt reflété dans le programme.

3.5 Bases de données réparties/machines bases de données

Les bases de données réparties constituent l'un des sujets les plus d'actualité en matière de techniques liées aux systèmes de gestion de bases de données. Dans le contexte d'une base de données réparties, les données s'étendent à plusieurs bases de données différentes (souvent sur des machines se trouvant à des endroits différents). Il incombe au logiciel du SGBD d'intercepter l'interrogation d'un utilisateur, de la traduire en interrogations appropriées aux diverses bases de données en présence et de grouper aux fins de présentation les résultats de ces interrogations.

données sans pour autant modifier l'interrogation. Cette dernière peut bénéficier sans délai de toute amélioration apportée à la structure de la base de données ou aux algorithmes d'optimisation.

De même, lorsque l'utilisateur formule une interrogation en SQL, il ne précise pas l'ordre suivant lequel doit se dérouler le traitement pour y répondre. Ce rôle incombe aux algorithmes d'optimisation du logiciel de traitement des interrogations. Ce logiciel évalue l'interrogation en fonction de la structure et de l'organisation actuelles de la base de données de manière à déterminer la façon la plus efficiente d'y répondre.

3.2 Norme universelle reconnue à l'échelle internationale

L'Organisation internationale de normalisation (ISO) et l'American National Standards Institute (ANSI) ont récemment adopté une norme commune à l'égard de SQL (ISO 1987). L'existence de cette norme et son adoption par plusieurs fournisseurs de systèmes de gestion de bases de données relationnelles permettent aux réalisateurs de logiciels d'avoir accès à un marché beaucoup plus vaste sans consentir des efforts supplémentaires considérables aux fins de la réalisation de ces logiciels. Comme ils fondent leurs applications sur le SQL normalisé, ils ne sont plus assujettis à un système particulier de gestion de bases de données. Par conséquent, la création d'outils logiciels en fonction d'une interface faisant appel à cette version normalisée de SQL constitue maintenant un secteur d'activité en pleine croissance. Par exemple, les interfaces faisant appel à un langage naturel, les langages de programmation de la quatrième génération, les logiciels de prise en charge des répertoires de données, les logiciels d'entrée/de validation de données, qui s'appuient tous sur le SQL adopté par l'ANSI et l'ISO, comment à apparaître sur le marché.

Le vif intérêt que suscite SQL a en outre eu une incidence très positive sur la norme elle-même et on continue de l'améliorer. Dans la version révisée provisoire la plus récente de la norme de l'ISO portant sur SQL, les contraintes liées à l'«intégrité référentielle» sont précisées dans les instructions SQL de définition des données. Pour bien illustrer cette extension de SQL, examinons de façon plus approfondie l'exemple «LOGEMENT». Si l'on émet comme hypothèse que la base de données comporte aussi un tableau intitulé «PERSONNES» faisant état de renseignements détaillés au sujet de particuliers, y compris un code de logement qui indique le logement où ils habitent à l'heure actuelle. On pourrait déterminer qu'il existe une contrainte d'intégrité selon laquelle chaque personne doit être associée à un logement en particulier. De ce fait, on commettrait une erreur en supprimant un enregistrement «LOGEMENT» auquel se rapporteraient encore des enregistrements «PERSONNES», ou en ajoutant un enregistrement «PERSONNES» qui se rapporterait à un enregistrement «LOGEMENT» non existant. A l'heure actuelle, la logique visant à repérer et à éviter ces incohérences doit être insérée dans chaque programme d'application qui peut supprimer un enregistrement «LOGEMENT», ce qui ne sera plus nécessaire si l'on précise dans SQL les contraintes d'intégrité référentielle. Le SGBD est chargé de repérer et de bloquer toute tentative de suppression d'un enregistrement «LOGEMENT» auquel contiennent de se rapporter des enregistrements «PERSONNES».

3.3 Facilité d'extension

Une des grandes différences entre les versions de SQL des divers fournisseurs réside dans le nombre et la diversité des fonctions prises en charge. Ces écarts sont en grande partie attribuables à la facilité avec laquelle il est possible d'intégrer à SQL des fonctions supplémentaires sans en modifier la structure d'ensemble. Par exemple, la norme SQL traite des fonctions de groupage aux fins de l'établissement d'une moyenne (AVG), de la détermination d'un nombre maximal (MAX) ou minimal (MIN), du comptage (COUNT) et de l'agrégation (SUM) à l'égard de données non pondérées. Si l'on revient à l'exemple «LOGEMENT», il serait possible d'obtenir au sujet du nombre d'occupants diverses statistiques sommaires réparties selon le lieu géographique:

Pour comprendre les possibilités qu'offre SQL, il est essentiel de se rendre compte que ce langage permet d'exécuter exactement ces quatre fonctions – rien de plus et rien de moins. Toute autre fonction doit être exécutée par l'application qui lance l'instruction SQL.

Examinons l'exemple suivant. Le tableau «LOGEMENT» fait état des données sur les logements, par exemple le nombre d'occupants, le type de logement, son emplacement, le genre de chauffage et l'âge du logement. Aux fins d'imputation du type de logement, on pourrait souhaiter obtenir un ensemble de logements constituant des données possibles qui sont situés dans la même région géographique, ont le même âge et font appel au même genre de chauffage. On pourrait présenter l'instruction SQL suivante en vue d'obtenir un ensemble de donneurs:

```
SELECT LOGEMENT_ID, TYPE_DE_LOGEMENT
FROM LOGEMENT
WHERE TYPE_OF_CHAUFFAGE = 'GAZ' AND
      AGE = 20 AND
      ENDROIT_CODE = 'XYZ';
```

(Interrogation 1)

SQL ne constitue pas un mécanisme dont on peut se servir pour manipuler l'ensemble d'enregistrements donneurs extrants. Ce langage ne permet pas de sélectionner le nième enregistrement, un enregistrement sur deux ou un enregistrement au hasard. De même, SQL ne possède aucun mécanisme qui permette de manipuler un tableau de manière à en modifier la présentation sur un terminal ou à l'impression. Ce sont là des possibilités qu'on exigerait à juste titre d'un langage de «programmation», d'où l'expression «langage d'interrogation de bases de données». Le fait de désigner SQL comme un langage de la quatrième génération, puis de le comparer à des produits qui mettent en jeu l'intégration à un langage de programmation des seules fonctions d'extraction et de modification de données ne fait qu'ajouter à la confusion. On établit alors une comparaison boiteuse; en effet, bien qu'il s'agisse dans les deux cas de langages de la quatrième génération, chacun présente des caractéristiques très différentes.

Compte tenu du nombre très restreint de fonctions qu'il permet d'exécuter, la question qui saute aux yeux est : «Alors pourquoi SQL suscite-t-il tant d'intérêt?»

3. SQL - LES AVANTAGES QU'IL PRÉSENTE

3.1 Transparence au moment de la mise en oeuvre

- quelles colonnes seront indexées (caractéristique contribuant à l'amélioration de la performance);
 - si le tableau ou la colonne est réellement stocké ou s'il s'agit uniquement de la combinaison, pour la durée d'exécution, d'autres tableaux; et
 - quelle est la représentation interne des données (c'est-à-dire la virgule flottante, décimal condensé, code binaire)
- n'ont aucun rapport que ce soit avec la syntaxe d'une instruction SQL. De ce fait, les modifications apportées à l'organisation et à la structure de la base de données n'ont aucune incidence sur l'utilisateur. On peut donc modifier selon son gré la structure sous-jacente à la base de

Note d'information sur SQL

DAVID N. EMERY¹

RÉSUMÉ

Cette note d'information met en lumière les points forts et les points faibles du langage SQL. MOTS CLÉS: Système de gestion de bases de données relationnelles; langage d'interrogation de bases de données.

1. INTRODUCTION

Les systèmes de gestion de bases de données relationnelles et SQL, le plus populaire des langages d'interrogation de bases de données, ont largement retenu l'attention des médias. Dans une large mesure, on voit en SQL un moyen de régler tous les problèmes liés à la gestion des données, ce qui entraîne malheureusement une grande confusion chez les utilisateurs possibles de ce langage. Il arrive donc que ces personnes soient déçues lorsqu'elles ont l'occasion d'utiliser SQL.

La présente note d'information a pour objet d'éliminer en partie la confusion qui existe dans ce domaine en traitant de SQL un portrait réaliste qui met en lumière les points forts et les points faibles inhérents à ce langage. Nous ne tenterons pas ici d'exposer les avantages que présente en soi le modèle de données relationnelles, car d'autres documents en ont déjà traité de façon appropriée (par exemple Date 1985).

2. SQL - LA NATURE DE CE LANGAGE

L'interaction qui s'exerce entre un utilisateur (qu'il s'agisse du réalisateur d'un système ou de l'utilisateur final) et le système de gestion d'une base de données peut se diviser en grandes catégories selon la fonction exécutée:

- définition de données;
- contrôle de données (c'est-à-dire autorisation et contrôle de leur intégrité);
- extraction de données; et
- modification de données (c'est-à-dire insertion, mise à jour et suppression).

Un système de gestion de bases de données doit fournir des interfaces aux fins de l'exécution de ces fonctions. Selon le système choisi, ces interfaces revêtent la forme de programmes utilitaires, de langages d'interrogation et/ou de bibliothèques de sous-programmes pour les langages de programmation.

Avec SQL, on exécute ces quatre fonctions à l'aide d'un seul langage, qui est bien défini et présente une structure rigide. SQL constitue l'interface dont on se sert pour communiquer, au système de gestion de la base de données, la façon dont les relations (c'est-à-dire les fichiers ou les tableaux logiques) doivent être subdivisées et/ou combinées de sorte que soient créées de nouvelles relations.

¹ David N. Emery, Statistique Canada, Sous-division de la recherche et des systèmes généraux, DSDI - bureau 2405, immeuble Principal, Ottawa, K1A 0T6.

5. CONCLUSION

Les résultats de l'analyse des adresses postales réalisée par le système PAAS sont encourageants. Le système décode la grande majorité des adresses, il génère un code très détaillé pour chaque composante, il normalise les composantes et produit une clé de recherche d'adresse de façon appropriée et il est capable de traiter les cas d'ambiguïté. En outre, le système PAAS comprend des programmes utilitaires ainsi que des interfaces pour les utilisateurs et les techniciens d'entretien.

Les utilisateurs ont accès à une interface, qui leur permet d'exécuter les quatre fonctions de base sur leurs adresses, ainsi qu'à un programme utilitaire qui traite les adresses erronées (traitement en direct). Ils disposent également d'un programme de traitement des fichiers. Le système PAAS est également pourvu d'un sous-ensemble de contrôle de la qualité destiné aux techniciens d'entretien. Le système PAAS est appelé à évoluer à l'avenir au fur et à mesure qu'on découvrira de nouvelles adresses et que d'autres adresses deviendront désuètes. Or, il est délicat de faire en sorte que les modifications appropriées soient apportées au système. Le sous-ensemble d'entretien a pour objet d'éviter que l'apport de modifications au logiciel n'invalide des adresses analysées de façon appropriée dans les versions précédentes du système.

REMERCIEMENTS

L'auteur tient à remercier J. P. Lozano et M. Viends, qui ont travaillé à la mise en oeuvre du système PAAS, ainsi que B. E. Hill et M. Elsaesser, pour leur participation à la rédaction du présent document. Enfin, il voudrait également exprimer sa gratitude à J. Armstrong pour avoir mis le système PAAS à l'essai et avoir formulé des commentaires à l'égard d'une version antérieure du présent article.

BIBLIOGRAPHIE

- BARRETT, WILLIAM A., BATES, RODNEY M., GUSTAFSON, DAVID A., et COUCH, JOHN D. (1986). *Compiler Construction*. Science Research Associates Inc.
- DEGUIRE, Y. (1987). Research into the parsing and standardization of free format addresses at Statistics Canada. Rapport interne, Statistique Canada.
- DREW, J. DOUGLAS, ARMSTRONG, JOHN, VAN BAREN, ALEX, et DEGUIRE, YVES (1987). Méthodologie de la construction d'un registre d'adresses à partir de plusieurs sources administratives. Symposium international sur l'utilisation des données administratives à des fins statistiques, Ottawa.
- HILL, TED (1986). *MPL A Translator Writing System*. System Documentation, 1-4. Statistique Canada.
- LOZANO, J. P. (1987). Postal Address Analysis System Study. Rapport interne, Statistique Canada.
- SOCIÉTÉ CANADIENNE DES POSTES (1986). *Répertoire des codes postaux: régions de l'Atlantique, du Québec, de l'Ontario et de l'Ouest*.
- STATISTIQUE CANADA (1986). Record Linkage Software User Guide. System documentation, Recherche et systèmes généraux.
- STATISTIQUE CANADA (1988). Postal address analysis system (PAAS): Project charter (ébauche). Rapport interne, Recherche et systèmes généraux.

Bien sûr, cette méthode permet uniquement de traiter les problèmes liés aux noms de municipalités. Ce qui n'est pas si mal, puisque ces problèmes représentent une proportion appréciable des cas d'ambiguïté et que ces cas sont faciles à déceler et à résoudre (ils ne nécessitent pas le traitement d'un grand nombre de données). On pourrait, dans le cadre de travaux ultérieurs, étudier l'opportunité de déceler et de résoudre d'autres cas d'ambiguïté. Enfin, quels que soient les perfectionnements apportés au logiciel, les adresses insolubles et non existantes continueront de poser un problème et il faudra assurer un suivi manuel à leur égard.

Adresse à analyser: 100 Rideau st Ottawa Ont K1N 5X2
À un certain stade, nous obtenons une chaîne de fragments d'adresse qui sera transformée par cinq règles. Le " | " indique un "OU" et [] est un élément syntaxique facultatif.

<NUMBER> <WORD> <ST_DESIGNATOR> <MUNICIPALITY> <PROVINCE> <PC>
RULE (1) <NAME> ::= <WORD | NAME> [WORD]
Chaîne de fragments d'adresse qui sera transformés par la règle (1).

<NUMBER> <NAME> <ST_DESIGNATOR> <MUNICIPALITY> <PROVINCE> <PC>
Nouvelle chaîne de fragments d'adresse générée par la règle (1).
Cette chaîne sera transformée par la règle (2). On notera qu'il serait appropriée qu'une opération sémantique introduite par la règle (2) détermine la composante du nom de la rue.

<ST_NAME> ::= <NAME>
RULE (2) <NUMBER> <ST_NAME> <ST_DESIGNATOR> <MUNICIPALITY> <PROVINCE> <PC>
<ST_NUMBER> ::= <NUMBER>
RULE (3)

<ST_NUMBER> <ST_NAME> <ST_DESIGNATOR> <MUNICIPALITY> <PROVINCE> <PC>
RULE (4) <ST_ADDRESS> ::= <ST_NUMBER> <ST_NAME> <ST_DESIGNATOR>
<ST_ADDRESS> <MUNICIPALITY> <PROVINCE> <PC>

<ADDRESS> ::= <ST_ADDRESS> <MUNICIPALITY> <PROVINCE> <PC>
RULE (5)

Le processus est terminé puisque la chaîne a été analysée au complet.

Figure 5: Règles pour un exemple de syntaxe d'adresse.

4.2 Normalisation des composantes

La normalisation a pour objet d'éliminer toute variation de nature stylistique des composantes définies au stade de l'analyse syntaxique.

Contrairement au programme ASKGEN2, le système PAAS ne tronque aucune composante et il retient tous les renseignements contenus dans chaque composante. Fondamentalement, cette normalisation s'effectue de trois façons différentes selon la nature de la composante.

1. COMPOSANTES POUVANT ÊTRE CODÉES

Toute composante pour laquelle existe un nombre limité de valeurs est normalisée en remplaçant la valeur de la composante par le code de CLASS de cette dernière (ce code identifie d'une manière unique la valeur normalisée de la composante). Font partie de la présente catégorie les composantes comme le nom de la province, la composante désignative de rue, etc.

2. COMPOSANTE DE NOM NON NUMÉRIQUE

Pour normaliser une composante de nom non numérique, il faut appliquer plusieurs règles afin de transformer la valeur initiale en valeur normalisée. Ces règles peuvent varier de l'élimination des caractères inutiles (par exemple guillemets, trait d'union, etc.) au remplacement des abréviations (par exemple Mtl devient Montréal).

3. COMPOSANTE DE NOM NUMÉRIQUE

Les composantes de nom numériques sont normalisées en exprimant le nom sous forme de chiffre. Par exemple Première devient 1, Deuxième devient 2, etc.

Ask

La clé de recherche d'adresse doit être brève et unique.

On s'assure du caractère unique de la clé en enchaînant dans un ordre prédéterminé les composantes normalisées d'une adresse (plutôt qu'à l'aide d'une table comme avec le programme ASKGEN2). Il convient de souligner que la clé de recherche d'adresse ne constitue pas nécessairement un identificateur unique des logements. Ainsi, dans les régions rurales, il arrive souvent qu'une même adresse postale représente de nombreux logements (par exemple RR #1 Ottawa Ontario).

Différentes techniques de condensation peuvent être mises en oeuvre pour abréger la clé. Toutefois, la condensation des données nécessite un certain temps et nous devons de choisir une technique efficace. Nous mettons actuellement à l'essai deux techniques différentes.

1. LA TRONCATURE

La présente technique consiste à tronquer les composantes de nom. Il ne s'agit pas d'une condensation réelle et elle pourrait modifier le caractère unique de la clé. Toutefois, elle constitue une technique simple et rapide.

2. LA CONDENSATION RÉELLE

La technique de condensation que nous étudions actuellement consiste fondamentalement à remplacer des combinaisons communes de caractères par un code de caractère non utilisé aux fins de la rédaction des adresses. Cette technique nous permettrait de préserver le caractère unique de la clé, mais elle rendrait sa production et son utilisation plus complexes. En conséquence, on prévoit qu'elle nécessiterait un temps de calcul plus long.

4.3 Résolution des ambiguïtés

Lorsqu'une ambiguïté a été décelée dans le cadre de l'analyse syntaxique, elle doit être résolue, manuellement ou automatiquement, par le système PAAS. À cette fin, le système PAAS utilise un fichier des noms de municipalités (ce fichier, élaboré à partir de la bande du répertoire de codes postaux de la Société canadienne des postes, comprend environ 6,000 noms de municipalités et couvre tout le pays).

spécifications de traduction et il a été utilisé à Statistique Canada aux fins de la mise en oeuvre de STATPAK (système d'extraction et de totalisation pour le recensement), NYSIS (programme de codage des noms) et NAMEPARS (programme d'analyse syntaxique des noms). Son utilisation permet de raccourcir le délai d'élaboration (ainsi, il n'est pas nécessaire de rédiger un programme personnalisé détaillé dans un langage de programmation classique comme le PL/1). De plus, les spécifications rédigées en BNF sont beaucoup plus faciles à comprendre qu'un programme dont la logique est complexe.

Le programme d'analyse syntaxique du PAAS exécute une analyse syntaxique assez complexe et il représente une application très importante du MPL. Ainsi, l'exploration des adresses s'effectue à l'aide d'un dictionnaire contenant plus de 600 termes, tandis que la mise en oeuvre de l'analyse syntaxique fait appel à plus de 100 règles de syntaxe. Dans le cadre de cette analyse, les composantes primaires initiales sont transformées selon une règle en vertu de laquelle les symboles figurant du côté droit sont définis par les symboles figurant du côté gauche et deviennent des segments d'adresses d'un niveau plus élevé (ce processus est appelé "enchaînement aval") jusqu'à ce que l'adresse ait été analysée au complet. Au cours de ce processus, une règle introduit une opération en vertu de laquelle les composantes d'adresses sont dégagées et stockées dans une table. Les adresses invalides sont décelées par le fait qu'aucune règle ne s'applique. On trouve à la figure 5 un exemple d'ensemble de règles. Enfin, certaines adresses complexes sont soumises à une analyse spéciale par l'introduction d'une opération sémantique au moyen du langage MPL. Une telle analyse est effectuée chaque fois qu'un terme ambigu est décelé. Dans un tel cas, le système PAAS analyse les composantes entourant le terme ambigu.

Le programme d'analyse syntaxique du PAAS offre un meilleur rendement que le programme ENCODA à divers titres.

- La qualité de l'analyse syntaxique - Le taux de réussite du programme d'analyse syntaxique du PAAS en matière de décodage est plus élevé que celui du programme ENCODA. Une série d'essais parallèles portant sur des échantillons nationaux d'adresses identiques ont démontré que le PAAS décode avec succès plus de 97% des adresses, tandis que le programme ENCODA ne peut en traiter correctement plus de 85%.

- L'indication de l'état de l'adresse - Le PAAS fournit un état plus complet de l'adresse que le programme ENCODA. En effet, ce dernier n'offre que deux possibilités: adresse décodée ou adresse en blanc!

- Les composantes - Le système PAAS génère des données beaucoup plus détaillées sur les composantes que le programme ENCODA.

- L'entretien - Grâce à l'utilisation du langage MPL, le programme d'analyse syntaxique du PAAS est d'un entretien beaucoup plus facile qu'un immense algorithme comme celui utilisé par le programme ENCODA.

ADRESSE	COMPOSANTE	TYPE	CAT	CLASS	AST_AMB
(1) 32 Main st, Ottawa, Ont	32	ST	NU	**	
	Main	ST	NA	**	
	st	ST	DE	**	
ADRESSE_ÉTAT===== > V	Ottawa	MU	NA	**	
	Ont	PR	NA	35	
(2) 32 Main st Ottawa Ont	32	ST	NU	**	
	Main	ST	NA	**	
	st	ST	DE	**	*
ADRESSE_ÉTAT===== > A	Ottawa	MU	NA	**	
	Ont	PR	NA	35	*

Figure 4: Exemples de résultats du traitement de l'adresse par le programme d'analyse syntaxique du PAAS.

Comme la deuxième adresse donnée en exemple n'est délimitée par aucune virgule, le PAAS signale une ambiguïté au moyen d'astérisques.

- L'entretien du logiciel constitue un véritable cauchemar. Les nouvelles structures d'adresses sont difficiles à intégrer aux programmes parce que ces derniers sont complexes et tendent à le devenir de plus en plus. Il s'agit là d'un signe certain d'obsolescence.

Afin de répondre aux besoins afférents à l'élaboration d'un registre d'adresses et aux besoins de l'analyse des adresses postales, on a donc entrepris l'élaboration d'un système complètement nouveau. Cette fois, l'approche adoptée pour solutionner le problème de l'analyse des adresses postales a fait appel aux techniques des systèmes experts, à la construction modulaire et à une mise en oeuvre à grande échelle. Le nouveau système porte le nom de système d'analyse des adresses postales ou PAAS.

4. UN SYSTÈME D'ANALYSE DES ADRESSES POSTALES: PAAS

Comme le système PAAS est actuellement en voie d'élaboration, certains des résultats obtenus sont strictement préliminaires, mais ils sont en général très encourageants. Nous allons passer en revue ci-après les quatre fonctions de base du système.

4.1 Analyse syntaxique des adresses

La fonction d'analyse syntaxique est la plus importante et la plus complexe des fonctions exécutées par le PAAS. Dans le cadre de son exécution, le système reçoit comme données d'entrée une adresse exprimée selon une structure non imposée, il explore cette adresse (la découpe en éléments lexicaux) et en effectue une analyse syntaxique pour la décomposer en composantes d'adresses. Pour chaque adresse traitée, le programme d'analyse syntaxique génère les éléments suivants (on trouvera deux exemples des résultats de ce traitement à la figure 4).

- Un code d'état de l'adresse très détaillé; comme V pour valide, E pour erreur de syntaxe, etc.
- Détermination des composantes de l'adresse d'entrée.
- Classification des composantes: chaque composante est classée à l'aide d'un code détaillé pour qu'il soit aisé d'en comprendre le sens. Ce code se subdivise en trois codes secondaires.
- Le code de TYPE qui indique à quel groupe de composantes une composante donnée appartient. On peut donner comme exemples de TYPES la province (PR), la municipalité (MU), la rue (ST), etc.
- Le code de CAT qui précise le groupe de composantes indiqué par le code de TYPE. On peut donner comme exemples de codes de CAT pour le code de TYPE rue (ST): le nom (NA), le numéro (NU), la composante désignative (DE), etc.
- Le code de CLASS qui permet de classer une composante selon ses caractéristiques. On peut donner comme exemples de codes de CAT pour une composante désignative de rue: l'avenue (AV) ou la route (RD).

- Détection des ambiguïtés: le programme d'analyse syntaxique du PAAS signale au moyen d'un astérisque toute composante susceptible d'être modifiée par suite d'une ambiguïté. Le programme d'analyse syntaxique du PAAS a été mis en oeuvre au moyen du langage MPL. Il s'agit d'un métalangage qui permet de produire des programmes ou des sous-programmes d'analyse syntaxique et de traduction automatique. Les données introduites en MPL sont un ensemble de spécifications portant respectivement sur l'exploration (la reconnaissance des composantes primaires), les règles de syntaxe et la sémantique. L'exploration consiste en une analyse lexicographique dans le cadre de laquelle l'adresse introduite est décomposée en composantes primaires. La spécification relative à la syntaxe est semblable à une spécification grammaticale en BNF: les symboles figurant du côté droit de la règle de syntaxe sont définis par les symboles figurant du côté gauche. On trouvera des exemples de règles de syntaxe à la figure 5. Enfin, toute règle peut introduire une opération sémantique utilisée pour traiter certains aspects complexes de la syntaxe ainsi que pour exécuter d'autres travaux (comme la mise à jour d'une table de composantes). Le langage MPL se prête bien à la rédaction des

Il peut s'agir de composantes d'adresses (normalisées ou non), ou encore des adresses initialement introduites, que ce soit pour des fins de suivi ou pour l'établissement d'un fichier historique. Toutefois, l'extraction des données d'une grande base de données (ou d'une grande table à deux dimensions) exige qu'on dispose d'une clé de recherche d'adresse (ASK) permettant tant d'avoir un accès direct à un enregistrement identifié par une adresse postale (ou permettant l'assortiment en direct d'un enregistrement et d'une adresse postale). Le traitement des étiquettes d'adresse représente un autre domaine où les adresses postales jouent un rôle de premier plan. En effet, les étiquettes d'adresse peuvent être formées à partir de composantes d'adresses normalisées ou non.

3.2 Trois éléments d'information de base

En conséquence, il faut pouvoir obtenir trois éléments d'information de base à partir d'une adresse exprimée selon une structure non imposée : les composantes d'adresses, les composantes normalisées et la clé de recherche d'adresse (ASK).

1. LES COMPOSANTES D'ADRESSES

Les composantes d'adresses représentent les parties reconnaissables et utiles d'une adresse. Les principales composantes d'une adresse sont le numéro de rue, le nom de la rue, l'orientation de la rue, la composante désignative de la rue, la composante désignative de l'entité postale, la composante qualitative de l'entité postale, le nom de la municipalité, le nom de la province et le code postal.

2. LES COMPOSANTES NORMALISÉES

Les composantes normalisées représentent la version normalisée des composantes d'adresses, toute variation d'ordre stylistique ayant été éliminée de ces dernières.

3. LA CLÉ DE RECHERCHE D'ADRESSE (ASK)

La clé de recherche d'adresse consiste en une chaîne de caractères condensée correspondant à une seule adresse.

3.3 Système d'analyse des adresses postales

Comme il remplace un spécialiste (comme un postier) aux fins de la reconnaissance des adresses, on peut considérer qu'un système complet d'analyse des adresses postales (à savoir un système informatique capable de produire les trois éléments d'information de base dont nous avons besoin) constitue un système expert dans le domaine du traitement des adresses postales. Au cours des années soixante-dix, deux programmes ont été élaborés à Statistique Canada pour analyser les adresses postales. Ces deux programmes, ENCODA (décomposition en composantes d'adresses) et ASKGEN 2 (normalisation et ASK), ont été mis en oeuvre pour être utilisés par le système de mise à jour du registre des entreprises. Jusqu'à récemment, ils ont bien rempli la fonction qui leur était dévolue. Toutefois, avec l'apparition d'ordinateurs plus puissants, de nouvelles techniques d'élaboration des logiciels et du registre d'adresses même, leur rendement ne satisfait plus aux normes actuelles.

Le taux de réussite des opérations de codage est trop faible. Une étude portant sur un échantillon national d'adresses tirées de nombreux fichiers administratifs démontre que, en moyenne, ENCODA est incapable de décoder 15% des adresses. Ce pourcentage d'erreur est inacceptable car, dans le cas de la création d'un registre national d'adresses, il pourrait entraîner plus d'un million d'erreurs de codage.

L'interface avec l'utilisateur est de piètre qualité. Aucun code d'état n'est produit au terme d'une analyse et très peu de programmes utilitaires ont été prévus pour faciliter la programmation. Les programmes ne sont pas pleinement fonctionnels. Ainsi, les composantes normalisées et la clé de recherche d'adresse font partie de la même structure de données. Les composantes normalisées sont tronquées pour permettre la condensation des données, mais la clé de recherche d'adresse est très longue car elle est stockée dans des zones de longueur fixe. En outre, le logiciel est incapable de déceler les adresses ambiguës.

2.2 La rédaction des adresses ne répond à aucune norme précise

Comme l'illustre la figure 2, une même adresse peut être rédigée de diverses façons. Cette situation s'explique par la souplesse de la syntaxe des adresses postales et par les caprices de la nature humaine. De fait, les gens écrivent les adresses comme bon leur semble et se contentent à cet égard aux "normes" en usage dans leur milieu.

2.3 Les adresses postales sont sources d'ambiguïtés

On ne peut étudier l'adresse postale uniquement d'un point de vue syntaxique, il faut aussi examiner son aspect sémantique (c'est-à-dire sa signification). Il arrive parfois qu'une même adresse puisse représenter divers lieux physiques. Nous sommes alors en présence d'une adresse ambiguë, que nous ne savons pas comment interpréter. À cette fin, il nous faudrait disposer de plus de renseignements pour pouvoir exclure les endroits inexistantes et découvrir l'endroit correct. Toutefois, ces renseignements supplémentaires ne nous permettent pas toujours de réduire à une seule les possibilités d'interprétation d'une adresse; nous faisons alors face à une ambiguïté insoluble. On trouve à la figure 3 un exemple d'adresse ambiguë.

3. CONSIDÉRATIONS D'ORDRE INFORMATIQUE

Maintenant que nous saisissons mieux la nature des adresses postales et des problèmes que pose leur interprétation, nous allons faire porter notre attention sur l'utilisation des adresses postales dans les systèmes informatiques.

3.1 Applications informatiques nécessitant le traitement de données sur les adresses

De nombreuses applications informatiques nécessitent le traitement de données sur les adresses. Certains projets de couplage des enregistrements (comme l'élaboration d'un registre d'adresses) nécessitent qu'on établisse des liens entre des individus ou des logements en se fondant sur les adresses postales correspondantes. Essentiellement, les règles de couplage mises en oeuvre à cet égard sont exécutées sur des composantes d'adresses normalisées. Les bases de données et les fichiers informatiques où sont stockées les adresses postales sont nombreux. Ainsi, les données sur les adresses postales utilisées aux fins de l'élaboration d'un registre d'adresses doivent être stockées quelque part, que ce soit dans une table autonome à deux dimensions ou dans une sorte de base de données intégrée. Mais quelles sont les données stockées?

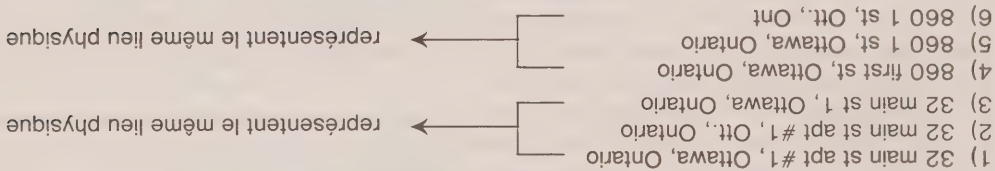


Figure 2: Exemples d'adresses représentant le même lieu physique.

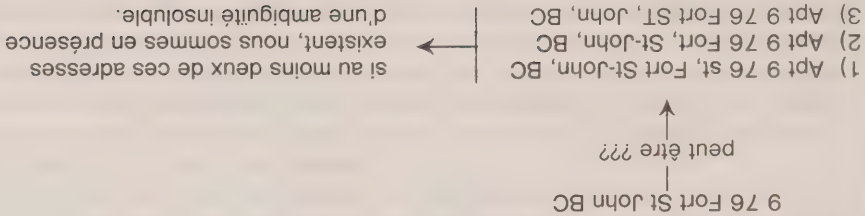


Figure 3: Exemple d'adresse ambiguë.

Comme on peut s'y attendre, la souplesse de la définition de l'adresse est source de problèmes pour toute application informatique nécessitant le traitement d'adresses postales. Même un être humain est susceptible d'éprouver certaines difficultés à décoder les adresses qu'il ne connaît pas bien. Nous analyserons ci-après les trois principaux problèmes qui se posent à cet égard.

2.1 La syntaxe des adresses postales canadiennes est complexe

Toute adresse postale est constituée de composantes primaires (éléments lexicaux qu'on peut considérer comme les unités de base de l'adresse). Une composante primaire peut être un délimiteur, un terme (ou mot clé), un mot, une lettre ou un nombre. On trouve à la figure 1 un exemple de décomposition d'une adresse en composantes primaires. Les composantes primaires peuvent être combinées pour obtenir des structures d'adresses plus importantes appelées 'composantes d'adresses'. Les composantes d'adresses peuvent elles aussi être de trois types : composantes désignatives, composantes qualificatives et mots secondaires. On trouve également à la figure 1 un exemple de décomposition d'une adresse en composantes d'adresses valides et d'un ensemble de combinaisons de composantes primaires valides. Toutefois, aux fins de la mise en oeuvre, il est plus pratique de définir une adresse à partir de combinaisons de composantes primaires. Ces combinaisons peuvent être produites à partir d'une grammaire formelle d'adresses postales (rédigée en BNF, par exemple) et être utilisées directement pour constituer une adresse postale.

La syntaxe des adresses postales est passablement complexe. Premièrement, la grammaire correspondante est assez volumineuse. Nous avons analysé un échantillon national de 30,000 adresses tirées de six fichiers administratifs différents de composantes primaires. Ces adresses environ 4,900 combinaisons différentes de composantes primaires. Ce chiffre est nettement plus élevé que celui déclaré par Drew (1987) : cet écart s'explique du fait que nous avons analysé des adresses provenant de nombreux fichiers différents, tandis que leur étude portait sur un seul fichier. Notre analyse nous a aussi permis de dégager d'autres résultats intéressants à l'égard de la répartition de ces combinaisons. Ainsi, il suffit de 37 combinaisons différentes pour pouvoir constituer 50% des adresses étudiées. Il existe donc un petit ensemble de combinaisons fréquentes, mais la plupart des combinaisons sont plutôt rares. Néanmoins, la présente analyse illustre bien la complexité de la syntaxe des adresses postales en démontrant qu'elle ne se limite pas à un petit nombre de combinaisons. Deuxièmement, on peut trouver jusqu'à 600 termes différents dans un bon échantillon national d'adresses. Troisièmement, les adresses sont d'ordinaire exprimées selon une structure non imposée, c'est-à-dire que les composantes (et les délimiteurs) peuvent figurer dans l'une ou l'autre d'un petit groupe de positions.

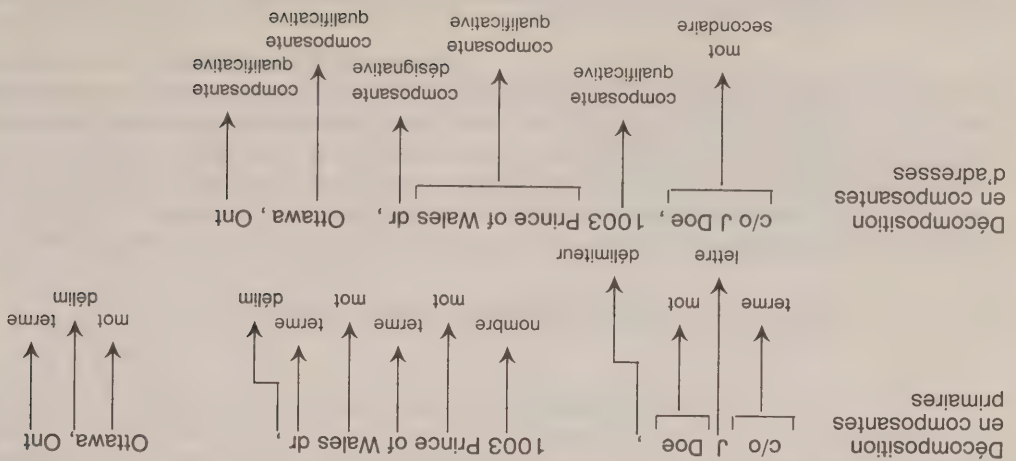


Figure 1: Deux façons de décomposer une adresse postale.

Analyse des adresses postales

YVES DEGUIRE¹

RÉSUMÉ

Lorsqu'on examine les adresses postales pouvant figurer dans un fichier administratif, nous sommes frappés par la complexité de la syntaxe, l'absence de normes, les ambiguïtés diverses et les nombreuses erreurs. L'utilisation des adresses postales par un système informatique présente donc de réelles difficultés. Le PAAS (système d'analyse des adresses postales) en voie d'élaboration à Statistique Canada a pour objet de remplacer le sous-programme désuet utilisé partout dans le Bureau pour décoder les adresses postales. Le PAAS permettra aux applications informatiques d'obtenir les composantes d'adresses, la version normalisée de ces composantes et la clé de recherche d'adresse (ASK) correspondante.

MOTS CLÉS: Adresses postales; données administratives; analyse syntaxique; normalisation; clé de recherche.

1. INTRODUCTION

On peut définir l'analyse des adresses postales comme le processus qui consiste à déterminer les composantes de base d'une adresse exprimée selon une structure non imposée, à normaliser ces composantes et à produire un identificateur pour cette adresse. Ce processus peut être mis en oeuvre, par exemple, à l'étape du traitement de toute application de couplage d'enregistrements utilisant une zone d'adresse ou encore pour produire une clé permettant de solliciter une base de données. Dans le cadre de son programme de recherche pour le recensement de 1991, Statistique Canada réalise une étude sur la mise en oeuvre d'un registre national d'adresses. Fondamentalement, un tel registre comprend des renseignements sur les adresses postales. Il faut analyser ces renseignements avec soin pour élaborer le registre et en évaluer la qualité. L'équipe de recherche sur le registre d'adresses a reconnu ce fait et entrepris des travaux de recherche dans le domaine de l'analyse automatisée des adresses postales.

Le présent mémoire fait état des résultats de cette recherche sur l'analyse des adresses postales. Nous y décrivons d'abord la nature de l'adresse postale et des problèmes que pose son interprétation. Nous examinerons ensuite certaines questions de nature informatique afin d'expliquer pourquoi nous avons besoin d'un nouveau logiciel pour le registre d'adresses et pour Statistique Canada. Enfin, nous étudierons le système PAAS, nouveau système d'analyse des adresses postales en cours d'élaboration à Statistique Canada.

2. ADRESSES POSTALES: ÉNONCÉ DU PROBLÈME

On peut définir l'adresse postale comme une chaîne de caractères représentant l'endroit où un individu peut prendre son courrier. Par endroit, on entend un lieu physique que le livreur (comme un postier) et le destinataire s'entendent à reconnaître comme lieu de réception du courrier. Il peut s'agir d'un logement, d'une case postale, d'une rue ou d'une route rurale. Afin de restreindre le champ d'observation de notre étude, nous allons examiner uniquement les adresses canadiennes (en langues française et anglaise) qui représentent des lieux de résidence et devaient permettre une livraison exacte du courrier.

¹ Yves DeGuire, Recherche et systèmes généraux, Statistique Canada, bureau 2405, immeuble Principal, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6.

BIBLIOGRAPHIE

- DODGE, H.F. (1950). Inspection for quality assurance. *Industrial Quality Control*, 7(1), 8.
- MacMILLAN, J.H., et MUDRYK, W.V. (1988). A non-parametric Empirical Bayes approach for estimating a process average in quality control. Communication présentée à la Section on Physical and Engineering Sciences, American Statistical Association Annual Meeting, New Orleans, Louisiana.
- MUDRYK, W.V., et BOUGIE, R.W. (1987). Quality Control Processing System (QCPs) - Users Manual. Document interne. Statistique Canada, Ottawa.
- SCHILLING, E.G. (1982). *Acceptance Sampling in Quality Control*. New York: Marcel Dekker.
- STATISTIQUE CANADA (1987). *Lignes directrices concernant la qualité*. 2^e édition, Ottawa, Canada.
- STEPHENS, K.S. (1982). *How to Perform Skip-Lot and Chain Sampling*. Volume 4, ASQC Basic References in Quality Control: Statistical Techniques. American Society for Quality Control, Milwaukee, Wisconsin.

- d'améliorer les aptitudes de l'opérateur en ce qui regarde le traitement;
- d'accroître sa motivation vis-à-vis de ses collègues;
- de le rendre plus sensible à la question de la qualité;
- d'améliorer son moral.

b. Avantages pour le superviseur

Le système fournit aux superviseurs l'information opérationnelle qui leur permet de mieux gérer le service dont ils sont responsables pour les questions suivantes:

- affectation efficace des ressources et répartition efficace du travail;
- identification des opérateurs ou des secteurs au sujet desquels il existe des problèmes;
- définition des besoins de formation.

c. Avantages pour la direction

Le système fournit à la direction des données sommaires sur des indicateurs fondamentaux du contrôle de qualité; grâce à ces données, la direction peut:

- s'assurer de la qualité du produit;
- suivre l'évolution du projet au point de vue de la qualité et des coûts;
- recommander une modification des objectifs opérationnels.

d. Avantages pour le concepteur des plans de CQ

Le système produit une information abondante (p. ex. estimations, séries chronologiques sur la qualité) qui sert à analyser le plan de contrôle de la qualité et à raffiner les méthodes et les procédures de chaque application. Lorsque le concepteur a recueilli des données de ce genre pendant un certain temps, il est en mesure:

- d'améliorer les méthodes et les procédures de CQ;
- d'apporter des corrections au plan d'échantillonnage ou à la méthode d'inspection;
- de réduire au minimum les frais d'inspection.

5. CONCLUSIONS

À Statistique Canada, le SCQ sert à appuyer les programmes de contrôle de qualité de nombreuses opérations d'enquête axées sur la production. Comme chaque programme a pour but ultime de prévenir les erreurs dans la mesure du possible et de réduire progressivement les proportions d'erreurs d'inspection à des contrôles par sondage, il est essentiel d'avoir un système de gestion souple et efficace. Le SCQ répond à ces objectifs puisqu'il produit rapidement des données exactes à l'intention du personnel des divers niveaux hiérarchiques affecté à chaque opération et sert de base aux diverses méthodes d'inspection qui s'inscrivent dans la stratégie globale du contrôle d'acceptation.

Le système est particulièrement intéressant pour les utilisateurs puisqu'il peut traiter facilement, rapidement et à peu de frais les opérations à grande échelle qui nécessitent de nombreux opérateurs. De plus, en considérant chaque opérateur individuellement, le système dirige l'attention vers chaque source d'erreur pertinente et produit à cette fin la rétroaction nécessaire pour les niveaux hiérarchiques appropriés. Ainsi, grâce à ce système, nos méthodes de contrôle de la qualité permettent de prévenir et, s'il y a lieu, de corriger les erreurs de façon efficace et économique.

REMERCIEMENTS

L'auteur tient à remercier les arbitres et Jeffrey Smith pour les commentaires constructifs qu'ils lui ont apportés lors de la révision de cet article. On peut obtenir des spécimens d'états imprimés produits par ce système en s'adressant à l'auteur.

- des listes sommaires d'indicateurs clés, qui permettent de faire une analyse avantages-coûts du CQ;
- une analyse de Pareto des effets de l'opérateur et du code d'erreur;
- des diagrammes de la qualité moyenne de la production des opérateurs par groupe, qui permettent de faire une analyse du rendement des opérations.

f. Rapports

Le système produit 8 rapports ordinaires et 5 rapports graphiques (par son lien avec SASGRAPH) pour chaque cas d'application. Il peut aussi produire des totalisations pour des sous-groupes déterminés (p. ex. les bureaux régionaux de Statistique Canada) ainsi qu'un résumé de chaque rapport pour l'ensemble des sous-groupes.

Chaque série de rapports est conçue à l'intention du personnel de quatre niveaux hiérarchiques: opérateur, superviseur, directeur et concepteur des plans de CQ. On peut se procurer des modèles de ces rapports en s'adressant à l'auteur.

3.2 Caractéristiques du logiciel

a. Capacité

Pour chaque application, le fichier du système peut contenir les dossiers de 108 opérateurs, chaque dossier pouvant contenir les données relatives à trois périodes antérieures. Le fichier a une caractéristique particulière: il élimine automatiquement les opérateurs qui ont été inactifs durant au moins un des quatre derniers mois consécutifs de traitement. Cette opération a pour effet de créer de l'espace dans le fichier pour de nouveaux opérateurs et, par le fait même, d'accroître la capacité effective du fichier.

b. Mise à jour de la série sur la qualité de la production de l'opérateur

Des que de nouvelles données sont obtenues, le système met à jour les séries sur la qualité de la production des opérateurs (jusqu'à 4 périodes consécutives). Il sera bientôt question de 6 périodes consécutives. Si un opérateur n'a pas traité de données durant un mois particulier, le système laisse un blanc pour ce mois. De même, le total cumulatif de l'année et le total trimestriel pour chaque application sont mis à jour dès que les données d'un nouveau mois sur le CQ sont recueillies.

c. Report de fin d'année

La plupart des applications du SCQ sont enregistrées en fonction de l'année civile. Quand c'est le cas, le système met à zéro les totaux mensuels antérieurs et crée une nouvelle série sur les applications (série qui débute habituellement en janvier). Toutefois, les totaux trimestriels et la série sur la qualité de la production de l'opérateur sont conservés.

d. Redressement

Si une totalisation est exécutée et que des erreurs sont constatées par la suite, la fonction "redressement" permet d'exécuter un nouveau passage machine avec les données corrigées de manière à obtenir automatiquement les résultats corrigés.

4. AVANTAGES DU SYSTÈME

Le SCQ a été créé pour répondre aux besoins de quatre niveaux hiérarchiques du personnel qui ont un lien avec chaque application de CQ. Par conséquent, les quatre rubriques suivantes sont celles qui conviennent le mieux pour décrire les principaux avantages de ce système.

a. Avantages pour l'opérateur

Le SCQ fournit à chaque opérateur une rétroaction complète sur son rendement passé et présent. L'opérateur est alors en mesure d'évaluer ses progrès, de comparer son rendement à celui de ses collègues et de savoir exactement où il commet des erreurs. Cette rétroaction a généralement pour conséquences:

c. Mesures de qualité pour les caractéristiques

Le système produira des estimations pour diverses mesures de la qualité, dont le pourcentage de valeurs erronées, le nombre d'erreurs par cent unités et des équivalents de l'erreur pondérée. En ce qui a trait à cette dernière mesure, le système prévoit une pondération des erreurs selon leur importance; il existe à cet effet un mode de classification préalable. De façon générale, on attribue aux erreurs un poids dont la valeur varie de 0 à 1, suivant l'importance relative de ces erreurs. Par souci de simplicité, on s'en tient habituellement à quatre catégories d'erreur:

Catégorie	Poids
Grave	1,0
Considérable	0,4 - 0,6
Minime	0,2 - 0,3
Négligeable	0,0 - 0,1

d. Estimations

Le système produit des estimations et les erreurs types correspondantes (s'il y a lieu) pour de nombreux indices fondamentaux du contrôle de la qualité, dont voici les plus importants:

(i) Taux d'erreur

Des taux d'erreur peuvent être calculés pour un opérateur, un plan d'échantillonnage ou l'application globale. Il existe des estimations pour diverses périodes de référence (p. ex. un jour, une semaine, un mois, un trimestre, etc.) et divers sous-ensembles de l'application comme des catégories de lot (p. ex. lots rejetés) ou des sous-groupes (p. ex. les bureaux régionaux).

(iii) Qualité moyenne de la production de l'opérateur

La qualité moyenne de la production de l'opérateur est une estimation du degré de compétence d'un opérateur à n'importe quel moment donné. On calcule cette estimation au moyen d'une méthode empirique de Bayes (MacMillan et Mudryk 1988), qui consiste essentiellement à rapprocher l'estimation courante du taux d'erreur de l'échantillon pour un opérateur du taux d'erreur global moyen des quatre dernières périodes pour cet opérateur. Cette opération est fondée sur le rapport entre la variance d'échantillonnage de l'estimation courante et la variance totale de l'estimation du taux global moyen. Il a été démontré que ce rapport produisait de bonnes estimations pour vérifier la compétence d'un opérateur pour des plans d'échantillonnage à inspection minimum.

(iii) Taux de rejet

Des taux de rejet réels et prévus sont calculés pour chaque plan d'échantillonnage dans le but de faire des comparaisons statistiques et d'évaluer les opérations. Pour calculer les taux prévus, on suppose des probabilités de Poisson.

(iv) Taux d'inspection

On calcule des taux d'inspection à divers niveaux pour avoir une idée générale des coûts relatifs. Des taux réels et prévus sont calculés en tenant compte ou non des effets de l'échantillonnage successif partiel. Les taux prévus sont un prolongement naturel des taux de rejet prévus mentionnés ci-dessus.

(v) Qualité moyenne à la sortie

Par suite de l'application du contrôle de la qualité à une opération, le système produit une estimation de la qualité moyenne à la sortie (QMS). Cette opération permet d'appliquer le taux d'erreur observé au niveau de l'opérateur au volume non inspecté pour ce même opérateur, puis de faire la somme pour tous les opérateurs afin de déterminer l'estimation global.

e. Analyse

Le système produit des totalisations et des résultats de traitement qui permettent de faire des analyses à divers niveaux et ensuite d'ajuster les paramètres d'application ou de modifier les plans. Ces résultats comprennent:

- des profils d'opérateurs permettant d'analyser l'utilité d'un plan ou d'une méthode d'échantillonnage;
- des évaluations de plans d'échantillonnage individuels, qui permettent de faire une analyse globale de plans de CQ;

- *Plans d'inspection gradués*. On obtient ces plans en augmentant ou en abaissant l'indice de qualité pour le plan d'échantillonnage dès qu'une variation de la qualité moyenne de la production est observée, puis en examinant de près les conséquences de ce rajustement pour les estimations de la qualité moyenne à la sortie.
 - *Plans de résultats cumules* ou, plus précisément, échantillonnage successif partiel (Stephens 1982). Dans ce cas, l'intervalle de prélèvement dépend de la stabilité et de la qualité prévue à l'entrée. Les deux méthodes font partie du contrôle d'acceptation et exigent une série de données chronologiques complètes qui indique non seulement la qualité du traitement (au niveau de l'opération) mais aussi le degré de stabilité (c.-à-d. de contrôle) auquel on peut s'attendre dans le processus. Par conséquent, l'inspection doit produire:
 - des données satisfaisantes (estimations d'erreur exactes);
 - des résultats rapides (mensuels, hebdomadaires, quotidiens);
 - des documents qui favorisent l'amélioration (rapports de contrôle – rétroaction);
 - des données chronologiques sur la qualité (chroniques sur le niveau d'erreur).
- Ce sont essentiellement ces considérations qui ont conduit à l'élaboration du Système de gestion de la qualité (SGQ). Il convient de souligner que le système est actuellement modifié pour élargir la série chronologique sur la qualité du travail des opérateurs. Cela devrait produire les données nécessaires pour nous permettre d'avoir plus souvent recours au contrôle par sondage ou au contrôle de processus pour certains opérateurs dont le rendement est stable ou exceptionnel.

3. DESCRIPTION DU SYSTÈME

- Compte tenu de la stratégie exposée ci-dessus, le SGQ a été élaboré en fonction des objectifs suivants:
- traiter toute opération d'échantillonnage pour acceptation simple;
 - produire des données pour chaque opérateur, celui-ci pouvant être considéré comme la source d'erreur;
 - procurer une rétroaction à quatre niveaux hiérarchiques du personnel par des données chronologiques et des données courantes sur le contrôle de la qualité;
 - appuyer le contrôle d'acceptation en permettant de traiter les résultats de l'échantillonnage successif partiel et en produisant des séries chronologiques complètes sur la qualité du travail des opérateurs;
 - soutenir les principaux objectifs du CQ, soit la correction et la prévention des erreurs, tout en permettant de réduire progressivement les frais d'inspection.

3.1 Caractéristiques méthodologiques

a. Plans d'inspection

Le système peut traiter n'importe quelle opération de contrôle de qualité découlant de l'application de l'échantillonnage pour acceptation simple. Cela comprend naturellement les plans normaux, réduits et renforcés de même que les lots qui n'ont pas été prélevés dans l'échantillonnage successif partiel. Le système traitera également les lots qui sont censés être inspectés entièrement.

b. Codes d'inspection du lot

Le système détermine le traitement des opérations de CQ par des codes d'inspection du lot, qui indiquent le degré d'achèvement de l'inspection. Il y a des codes d'inspection du lot pour les cas suivants:

- échantillon inspecté et accepté;
- échantillon inspecté et rejeté (reste du lot inspecté);
- lot inspecté entièrement;
- inspection non terminée (pour l'un ou l'autre des trois cas précédents);
- aucune inspection à cause de l'échantillonnage successif partiel.

2. STRATÉGIE RELATIVE AU CONTRÔLE DE LA QUALITÉ

2.1 Méthodes de contrôle de la qualité

Si nous considérons les deux principales méthodes de contrôle de la qualité, soit les cartes de contrôle du processus et l'échantillonnage pour acceptation, nous constatons que la seconde, appliquée selon les règles du contrôle d'acceptation, est celle qui se prête le mieux au contrôle *direct* de la qualité des opérations de traitement et ce, pour les raisons suivantes :

- on ne peut supposer au départ un contrôle préalable ou la stabilité du processus; en outre, l'une et l'autre condition ne sont pas toujours réalisées à long terme;
- on ne connaît pas toujours les causes d'erreur identifiées puisqu'il s'agit de personnes (plutôt que de machines);
- on ne peut interrompre facilement les processus pour en éliminer les causes d'erreur identifiées même si celles-ci sont connues;
- compte tenu du grand nombre d'opérateurs et de la forte variabilité d'erreur d'un opérateur à l'autre, il faudrait prévoir, parallèlement aux opérations d'enquête, un grand nombre de cartes de contrôle individuelles qui exigeraient une mise à jour immédiate (c.-à-d. après chaque observation de l'échantillon); cela serait difficilement réalisable sur le plan opérationnel.

Par conséquent, la stratégie de contrôle de la qualité consiste normalement à appliquer au niveau de l'opérateur diverses méthodes d'échantillonnage pour acceptation (avec correction), lesquelles constituent un processus d'élimination préliminaire visant à redresser les données de qualité inférieure; en dernier ressort, cette stratégie a pour but de réduire constamment le nombre d'inspections en se servant des résultats de ces inspections pour évaluer les progrès accomplis. Parallèlement, on met l'accent sur la rétroaction pour les opérateurs et les superviseurs en vue de prévenir les erreurs. On se trouve ainsi à corriger puis à prévenir les erreurs à la source, c'est-à-dire là où l'incidence de ces deux opérations (correction et prévention) peut être la plus forte. Par ailleurs, les variations d'un opérateur à l'autre sont traitées automatiquement puisque chaque opérateur est effectivement considéré comme un processus au sens suivant: dans une période de stabilité faible ou moyenne, on applique l'échantillonnage pour acceptation à chaque lot traité. Dans une période de forte stabilité allée à des résultats d'inspection favorables, on peut réduire l'échantillonnage pour acceptation et même le contrôle par sondage peut être mis en application dans le cadre plus large du contrôle d'acceptation.

2.2 Contrôle d'acceptation

Quand un programme de contrôle de la qualité est en application depuis un certain nombre d'années, l'opérateur tend à s'améliorer et, dans beaucoup de cas, le niveau de qualité se stabilise. Pour tirer profit de cette situation et rendre nos plans de contrôle de la qualité plus économiques, nous avons adopté la méthode que Schilling appelle le contrôle d'acceptation (1982). Selon cette méthode, les modalités de l'échantillonnage pour acceptation sont révisées continuellement en fonction des changements observés sur le plan de l'inspection. Cela rejoint l'opinion de l'un des principaux promoteurs du contrôle de la qualité, H. F. Dodge, qui affirme (1950): "Un bon produit qui a toujours été réputé être de bonne qualité n'exige pas autant d'inspection qu'un produit pour lequel on n'a aucune donnée particulière ou qui est réputé être de qualité irrégulière. Il est donc recommandé d'inclure dans les règles d'inspection des dispositions permettant de réduire ou d'accroître le nombre d'inspections suivant le genre et la quantité de renseignements dont on dispose au sujet du niveau de qualité et du degré de contrôle." (Traduction)

De fait, le but ultime du contrôle d'acceptation est de ramener progressivement les procédures d'inspection à des contrôles par sondage ou à des contrôles de processus à mesure que le niveau de qualité moyen augmente pour ensuite se stabiliser. Deux méthodes particulières sont utilisées à Statistique Canada pour atteindre cet objectif:

- planification
- élaboration
- réalisation
- traitement
- publication.

Il convient de souligner qu'il peut se produire des erreurs à chacune de ces étapes et que plus on met de temps à découvrir les erreurs dans le processus, plus les conséquences sont marquées pour ce qui a trait au coût de l'enquête, aux délais de production et à la précision des données. Il faut donc dès le début accorder toute l'importance voulue à l'élaboration de mesures et de méthodes visant à prévenir les erreurs ou à en réduire le nombre. Cela devrait se faire au moment de la planification et de l'élaboration de l'enquête. C'est ce qu'on appelle l'assurance de la qualité.

1.2 Assurance de la qualité

En matière d'assurance de la qualité des opérations d'enquête, il y a une approche générale qui consiste à prévoir les problèmes très tôt dans le processus et de prendre les mesures nécessaires pour les prévenir ou les atténuer. Ce travail préparatoire peut reposer sur l'expérience ou sur des analyses, des évaluations, des séances de compte rendu, des études de faisabilité, etc. Les mesures adoptées pourraient être l'amélioration des bases ou des plans de sondage, la modification des méthodes de collecte des données, l'amélioration des questionnaires, la clarification des modalités de traitement, etc. Le document de Statistique Canada intitulé Lignes directrices concernant la qualité (1987) renferme une liste complète de mesures conçues pour prévenir ou atténuer les problèmes.

Cette approche est essentielle car elle contribue effectivement à accroître le niveau de qualité en amont du processus et, de ce fait, empêche l'apparition de nombreux problèmes. En outre, elle garantit une qualité supérieure à un coût minimum en permettant d'obtenir des données justes dès le départ. Malgré toutes ces mesures, il y a encore des cas où les niveaux d'erreur sont inacceptables. Il faut alors envisager des mesures de contrôle de la qualité.

1.3 Contrôle de la qualité

Contrairement à l'AQ, le contrôle statistique de la qualité se fait surtout au moment du traitement. À cette étape, les tâches présentent habituellement les caractéristiques suivantes:

- répétitives et à fort coefficient de main-d'oeuvre;
- assignées à des personnes qui ont des aptitudes différentes;
- groupées normalement en fonction d'unités de travail semblables.

Les opérations correspondantes sont en soi plus exposées aux erreurs. Parmi ces opérations, notons:

- codage/transcription
- vérification manuelle/révision
- saisie des données/entrée
- corrections/rapprochement
- mise à jour/définition de profil, etc.

Pour plusieurs raisons, dont la complexité des tâches, la compétence des opérateurs, la rotation du personnel, etc., la fréquence et l'importance des erreurs varient d'une opération à l'autre, d'un opérateur à l'autre pour une même opération et, parfois, pour le même opérateur. Le contrôle statistique de la qualité vise à déterminer et à réduire cette variabilité et à faire en sorte que la qualité de chaque opération à la sortie soit d'un niveau acceptable.

Système de gestion de la qualité dans les opérations d'enquêtes¹

WALTER MUDRYK²

RÉSUMÉ

Les méthodes servant à contrôler la qualité des opérations d'enquête à Statistique Canada consistent habituellement en un échantillonnage pour acceptation pour chaque caractéristique, accompagné d'une inspection de redressement, l'une et l'autre opération s'inscrivant dans le cadre plus général du contrôle d'acceptation. Bien que ces méthodes soient considérées comme de bonnes mesures correctives, elles sont peu efficaces pour empêcher une répétition des erreurs. Vu l'importance primordiale de cet aspect de la gestion de la qualité, le Système de gestion de la qualité (SGQ) a été conçu en fonction de plusieurs objectifs dont la prévention des erreurs est un des principaux. C'est pourquoi il sert à produire des rapports de contrôle et des graphiques à l'intention des opérateurs, superviseurs et gestionnaires chargés des diverses opérations. Il sert également à produire des données sur les changements survenus au chapitre de l'inspection et permet ainsi aux méthodologistes de réviser les plans et les méthodes d'inspection en conformité avec les grandes lignes du contrôle d'acceptation. Cet article expose les principales caractéristiques du SGQ au point de vue de l'estimation et de la totalisation des données et montre de quelle façon ce système dessert les principaux programmes de contrôle de la qualité à Statistique Canada. Des fonctions importantes sont également analysées du point de vue méthodologique et systémique.

MOTS CLÉS: Système de gestion de la qualité; contrôle du processus; échantillonnage pour acceptation; contrôle d'acceptation; échantillonnage successif partiel.

1. INTRODUCTION

Cet article porte principalement sur les caractéristiques du Système de gestion de la qualité (SGQ) actuellement en usage à Statistique Canada. Cependant, avant de voir comment ce système sert au contrôle de la qualité dans les enquêtes, nous examinons brièvement le processus d'enquête et le rôle de l'assurance et du contrôle de la qualité dans ce processus. Nous voyons ensuite les méthodes de contrôle de la qualité appliquées particulièrement aux opérations de traitement à Statistique Canada et examinons le rôle du SGQ dans ce contexte. Enfin, nous donnons une description des caractéristiques et faisons un résumé des principaux avantages du système.

1.1 Processus d'enquête

Le maintien de la qualité dans le processus d'enquête a toujours été une préoccupation majeure à Statistique Canada. Dans un sens très général, on peut dire que le maintien de la qualité est assuré par l'application d'une série de mesures d'assurance de la qualité (AQ) et de contrôle de la qualité (CQ) à des étapes précises du processus. Il importe de faire la distinction entre ces deux activités puisque, à Statistique Canada, elles supposent des approches et des méthodes très différentes qui en règle générale ne sont pas appliquées aux mêmes étapes du processus. À Statistique Canada, le processus d'enquête comporte en gros les étapes suivantes:

¹ Cet article est une version révisée de la communication présentée à la quatrième conférence de recherche annuelle de Statistique Canada, 10-J, Immeuble Coats, Parc Tunney, Ottawa (Ontario), Canada, KIA 0T6.

² W. V. Mudryk, Division des méthodes d'enquêtes-entreprises, Secrétaire de l'informatique et de la méthodologie, du U.S. Bureau of the Census, Arlington (Virginie), mars 1988.

BIBLIOGRAPHIE

- APPEL, M., et HELLEBRMAN, E. (1983). Census bureau experiments with automated industry and occupation coding. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 32-40.
- CONNOR, J., SALLIUM, B., et WENZOWSKI, M. (1988). ACTR Documentation Set. (System Overview, User's Guide, Tutorial, Guide to the Parsing Strategy, Default Parsing Data, Message Guide, Command Language Guide, Searching and Matching Methods & Programmer's Guide) Documents internes, Statistique Canada, Sous-division de la recherche et des systèmes généraux, Ottawa, Canada.
- LANDRY, L., et PIDCOCK, J. (1984). Business Register Automated SIC Coding System, System Proposal and Design. Document interne, Statistique Canada, Division des services et du développement informatiques, Ottawa, Canada.
- REGHBATI, H. (1981). An overview of data compression techniques. *Computer*, 14,4, 71-75.

La formule utilisée pour le calcul des cotes optimistes est fondée sur les éléments suivants: a , le nombre de mots connus dans l'expression étudiée; b , le nombre de mots de l'expression étudiée qui ont déjà fait l'objet d'une recherche; c , le nombre total de mots dans l'expression étudiée; et d , le nombre de mots connus qui n'ont pas encore fait l'objet d'une recherche $a - b$; En conséquence, la formule de calcul des cotes optimistes est: $(d^2 \times \sum_{i=1}^a w_i) / c$.

Une fois la table des cotes optimistes établie, le système calcule la cote probable à chaque itération avant d'explorer la base de données. On se trouve ainsi à éliminer les recherches vaines. En résumé, la recherche d'une concordance indirecte se termine lorsque l'une ou l'autre des conditions suivantes est satisfaite:

1. La cote maximum probable pour l'itération courante est inférieure à la limite définie pour les concordances possibles.
2. Au moins une concordance a été établie et la cote maximum probable pour l'itération courante indique qu'aucune autre concordance ne peut être établie.
3. Le nombre maximum de concordances possibles que souhaitait l'utilisateur ont déjà été établies et la cote maximum probable pour l'itération courante est égale ou inférieure à la plus petite cote attribuée à une expression.

5. SOMMAIRE

Nous avons exposé dans cet article une méthode de codage automatique souple et efficace intégrée dans un système d'applications informatiques généralisé. Ce système permet le codage automatique de textes de langue anglaise ou française suivant n'importe quel mode de classification. À cette fin, il utilise une stratégie d'analyse généralisée puissante et des techniques d'optimisation de rendement élevé. Pour plus d'informations sur le ACTR, prière de consulter la source Connor, Salloum et Wenzowski (1988).

6. REMERCIEMENTS

Bien qu'il serait trop long de nommer toutes les personnes qui ont participé de près ou de loin au projet ACTR, l'auteur tient à remercier plus particulièrement les collaborateurs suivants: John Connor et Bill Salloum, les principaux programmeurs qui ont participé au projet; Victor Estevao, qui a contribué à l'élaboration de la stratégie d'analyse; Malvinder Rakhra et Paul Surman, qui ont soumis le système à de nombreux tests; et Don Royce, qui a été chef de projet.

suivants: n , le nombre de mots communs aux deux expressions; w_k , le poids du mot k ; m , le nombre de mots dans l'expression étudiée; et l , le nombre de mots dans l'expression de la base de données.

Par conséquent, la formule pour le calcul des cotes est: $n^3 \times \sum_{k=1}^n w_k/m \times 1$.

Paramètres d'appariement

Après avoir calculé une cote pour chaque concordance possible, le système compare cette cote à la valeur fixée par l'utilisateur pour les paramètres suivants et exécute l'opération indiquée.

1. LIMITE SUPÉRIEURE

Si la cote est égale ou supérieure à cette valeur, une concordance a pu être établie effectivement.

2. LIMITE INFÉRIEURE

Si la cote est égale ou supérieure à cette valeur mais moindre que la limite supérieure, une concordance est reconnue comme possible.

3. ÉCART EN POURCENTAGE

Si le système a pu établir plus d'une concordance et que l'écart entre les cotes concordantes est inférieur à la valeur fixée pour ce paramètre, on en déduit que plusieurs concordances ont pu être établies pour la même expression.

Limitation de la recherche pour un appariement indirect

Lorsque le système explore la base de données pour y dénicher des concordances possibles, il le fait à l'aide des mots connus de l'expression étudiée. Autrement dit, le système se sert des mots connus pour découvrir dans la base de données les expressions qui les renferment. La recherche s'effectue par ordre croissant de fréquence d'occurrence des mots connus. Ainsi, la recherche débute avec le mot connu qui revient le moins souvent dans la base de données, se poursuit avec le mot qui revient le moins souvent après le premier et ainsi de suite.

Comme on peut le constater, la recherche d'une concordance indirecte est une opération qui risque d'être fastidieuse et très coûteuse. Malheureusement, la recherche de concordances par voie indirecte est inévitable puisque l'appariement d'expressions très comparables est une caractéristique essentielle de tout système de codage automatique.

Au cours d'une opération de recherche, le système dresse une liste des expressions de la base de données qui ont été évaluées. Après qu'une expression a été évaluée, elle n'est pas ré-évaluée dans une itération subséquente du même exercice d'appariement. De cette façon, on est sûr qu'une expression qui renferme plus d'un mot connu ne sera pas évaluée une seconde fois. Dans le but d'optimiser davantage la recherche, le système utilise les paramètres d'appariement définis par l'utilisateur. À l'aide de ces paramètres, il construit une table de cotes optimistes pour chaque itération de la recherche axée sur le mot:

1. Pour le premier mot connu, la cote optimiste est fondée sur l'occurrence probable d'une expression de la base de données qui renfermerait un nombre de mots équivalant au nombre de mots connus et dont tous les mots correspondraient aux mots connus de l'expression étudiée.

2. Une hypothèse semblable est avancée pour le second mot; toutefois, comme le premier mot a déjà servi dans l'itération précédente, nous savons que les expressions qui renferment ce mot ont déjà été évaluées. Par conséquent, la cote optimiste est fondée uniquement sur la présence du second mot et des mots subséquents.

3. Les cotes optimistes pour des itérations successives reposent essentiellement sur la présence de mots qui n'ont pas encore fait l'objet d'une recherche pour l'itération courante et les itérations subséquentes.

clé d'extraction de données, qui est une forme d'utilisation efficace. La compression de la chaîne de caractères est régie par les règles suivantes:

1. En règle générale, les caractères qui forment les mots résultant de la stratégie d'analyse produisent uniquement de l'ensemble des 26 caractères alphabétiques et des 10 caractères numériques. (Il convient de rappeler que l'ensemble des caractères que l'on peut retrouver dans les mots est défini par l'utilisateur.) Par ailleurs, comme les caractères sont enregistrés (c.-à-d., en mémoire ou sur disque) à l'aide d'un code de 8 bits, il existe 2^8 ou 256 combinaisons possibles tandis qu'il n'en faut pas plus que 36 habituellement pour les mots traités par le ACTR. Il reste donc 220 codes qui peuvent servir à d'autre chose.
2. Certaines combinaisons de deux ou trois lettres reviennent plus souvent que d'autres dans les échantillons de textes anglais ou français. Dans le ACTR, les combinaisons de deux lettres sont appelées des «digrammes» et les combinaisons de trois lettres, des «trigrammes».
3. Les 220 codes «inutiles» peuvent alors servir à remplacer les digrammes et les trigrammes à mesure que ceux-ci sont relevés dans les échantillons de texte.
4. Une fois la stratégie d'analyse et l'enchaînement terminés, le ACTR parcourt la chaîne de caractères pour y repérer des digrammes et des trigrammes. S'il en trouve, il les remplace par le code de 8 bits correspondant. La conséquence directe de cette opération est qu'une suite de caractères qui exigeait auparavant 16 ou 24 bits de mémoire n'en exige plus que 8.

Appariement indirect

Comme l'appariement direct, l'appariement indirect est fondé sur des mots qui résultent de la stratégie d'analyse mais ne peut jamais être aussi efficace car la notion de concordance quasi-parfaite est relative, c'est-à-dire qu'il n'est pas possible de trouver l'expression la plus comparable sans examiner au préalable toutes les concordances possibles.

Pour exécuter un appariement indirect, il faut tout d'abord explorer la base de données pour appariement afin de vérifier si les mots ayant fait l'objet de la stratégie d'analyse ne s'y trouveraient pas. Ensuite, il faut extraire et évaluer toutes les expressions qui renferment chaque mot trouvé.

Le système détermine l'expression la plus comparable en calculant une cote pour chaque concordance possible. Les cotes sont établies en fonction des poids des mots qui se trouvent à la fois dans la base de données et dans l'expression étudiée. L'expression qui reçoit la plus haute cote est alors reconnue comme l'expression la plus comparable.

Calcul du poids d'un mot

Pour chaque mot identifié dans la base de données, le ACTR calcule une valeur heuristique d'appariement ou poids. Ces poids sont un indice de l'utilité d'un mot dans l'attribution d'un code et entrent dans le calcul des cotes pour les expressions.

La méthode de calcul des poids est fondée sur les éléments suivants: n , le nombre de codes individuels se rapportant aux expressions qui renferment le mot en question; V_i , la fréquence relative du code i selon des enquêtes antérieures; X_i , le nombre d'occurrences du mot pour les expressions portant le code i ; P_i , la proportion des occurrences de ce mot pour le code i , calculée comme suit: $V_i \times X_i / \sum_{j=1}^n V_j \times X_j$; EW , l'entropie du mot, calculée comme suit: $-\sum_{i=1}^n P_i \times \log_2 P_i$; K , le nombre total d'occurrences du mot pour le code i , calculé comme suit: $\sum_{i=1}^n X_i \times E_i$; EU , l'entropie d'une variable distribuée uniformément avec K valeurs individuelles, calculée comme suit: $\log_2(K)$; et enfin EO , une petite valeur permettant d'éviter une division par 0 et calculée comme suit: $-K/K + 1 \times \log_2 K/K + 1$. Par conséquent, la formule pour le calcul des poids est: $EU-EW + EO/EO + EW$.

Calcul de la cote d'une expression

Une cote est calculée pour chaque expression de la base de données qui fait l'objet d'une évaluation en vue d'un appariement indirect. La méthode de calcul est fondée sur les éléments

4. MÉTHODES DE RECHERCHE ET D'APPARIEMENT

Le ACTR soumet toujours le texte en question à la stratégie d'analyse avant d'exécuter un appariement. Si, une fois cette étape franchie, le ACTR trouve dans la base de données pour appariement un texte en tout point conforme au texte étudié, on parle alors de «appariement direct». Si une appariement direct ne peut être établie, le ACTR peut, à la discrétion de l'utilisateur, poursuivre l'exploration de la base de données pour trouver le texte le plus comparable. C'est ce qu'on appelle la «appariement indirect». Bien qu'elles soient fondées toutes deux sur des textes ayant fait l'objet d'une stratégie d'analyse, les deux méthodes d'appariement utilisées par le système ACTR diffèrent grandement l'une de l'autre par leur mécanisme d'identification des concordances.

Appariement direct

Dans l'appariement direct, on ne vise qu'à établir des concordances totales. Il convient de rappeler que l'appariement porte sur des textes ayant fait l'objet d'une stratégie d'analyse; par conséquent, les textes entre lesquels il existe une concordance totale peuvent sembler différents sous leur forme originale. Cela est une conséquence directe de la stratégie d'analyse utilisée. En ce qui regarde l'accès à la base de données, l'utilisation d'une clé est le moyen le plus rapide pour atteindre un élément de données. Malheureusement, pour pouvoir appliquer cette méthode aux expressions du ACTR dans leur forme originale, ces expressions ne doivent pas comporter plus de 200 caractères et plus de 20 mots ayant été soumis à la stratégie d'analyse. Ces deux exigences rendent irréalisable l'accès par clé puisque la longueur exceptionnelle de la clé contrebalancerait tous les avantages que l'on pourrait tirer de cette méthode. La seule autre solution à laquelle on peut songer est l'accès séquentiel mais cette solution n'est pas souhaitable à cause du temps qu'elle requiert pour examiner la grande quantité d'information que renferme d'ordinaire une base de données pour appariement.

Nous n'avons donc pas d'autre choix que réduire d'une quelconque façon la longueur de la clé, rendant ainsi possible l'accès par clé. Il existe de nombreuses méthodes de compression de données bien connues pour réaliser cela; on procède à la compression des données en créant la «clé d'expression condensée» ou CEC. Nous expliquons plus loin la façon de créer cette clé. Disons pour l'instant que la longueur de la CEC équivaut à environ 35% de la longueur initiale de l'expression. La CEC donne donc accès à la base de données pour appariement afin de déterminer l'existence de concordances directes.

1. Toutes les concordances totales seront toujours établies à l'aide de cette méthode.
2. Comme le ACTR permet d'établir des concordances directes le plus efficacement possible, les concordances établies de cette façon sont exécutées plus rapidement et à moindre coût.
3. À mesure que les applications se développent, la proportion des concordances directes s'accroît en règle générale à cause de la mise à jour soutenue de la base de données par l'utilisateur. Il peut donc y avoir une baisse réelle des coûts totaux d'appariement pour une application à mesure que celle-ci se développe, en dépit d'un élargissement de la base de données pour appariement.

Création d'une CEC

On crée une CEC en ordonnant tout d'abord les mots définis dans la stratégie d'analyse. L'ordre de classement est choisi arbitrairement et n'est donc pas important, dans la mesure où c'est le même pour toutes les CEC. (Il s'agit habituellement de l'ordre lexicographique croissant.)

Après avoir été ordonnés, les mots sont groupés en une seule chaîne sans espace. On compile ensuite cette chaîne de manière à la rendre suffisamment courte pour qu'elle serve de

pas avec «Take-Out» ou «Take-Out». Toutefois, si l'on définissait l'expression «Take-Out» avec comme substitut «Take-Out», on résoudreait la question. Nous avons ici un exemple qui montre comment combiner des étapes de la stratégie d'analyse. Si nous prenions aussi en considération l'exemple du mot à trait d'union, tous les cas de mots à trait d'union et de mots doubles ou simples conviendraient.

Mots sans importance: Si un tel mot est repéré en cours d'analyse, il est supprimé sans autre considération.

Par exemple, si l'ensemble des mots sans importance contient les éléments «A», «Am» et «I» et que l'on relève les deux expressions «I Am A Computer Programmer» et «Computer Programmer», il y aura concordance des deux expressions.

Suffixes: Cette fonction permet de scruter un mot de droite à gauche pour y trouver la plus longue forme de suffixe définie, de sorte qu'une fois le suffixe enlevé, le mot contienne au moins cinq caractères. Si une forme définie de suffixe est repérée, elle est supprimée.

Par exemple, si les suffixes «ing» et «er» sont définis, les expressions «Computer Programming» et «Computer Programmer» concorderont.

Remplacement de suffixes: Comme la précédente, cette fonction permet de scruter un mot de droite à gauche pour y trouver la plus longue forme de suffixe définie. Si un tel suffixe est repéré, il est remplacé par le substitut prévu.

Par exemple, un utilisateur pourrait vouloir éliminer d'un mot la marque du pluriel de telle manière que le suffixe du singulier soit reconnu à l'étape de la suppression des suffixes. Une illustration est fournie par les expressions «Battery Manufacturing» et «Manufacturing Batteries». Si le suffixe «ies» est remplacé par «y», non seulement ces expressions seront identiques, mais aussi elles subiront le même traitement à l'étape de la suppression des suffixes.

Consonnes ou voyelles doubles: À ce stade de l'opération, le système examine chaque mot pour y déceler des occurrences de consonnes ou de voyelles doubles appartenant à l'ensemble défini par l'utilisateur. Si une telle occurrence est repérée, le système élimine une des deux lettres. En règle générale, l'ensemble des consonnes et voyelles doubles comprend tous les caractères alphabétiques. Si tel est le cas, une concordance pourra être établie entre «Programmer» et «Programmer» malgré l'erreur d'orthographe.

Mots racine: Par cette fonction, le système examine les mots pour y déceler des mots racine. Le système examine le mot de gauche à droite et cherche la plus longue forme de mot racine définie. S'il en identifie une, le substitut remplace le mot entier et les étapes de la suppression et du remplacement des suffixes sont sautées.

Par exemple, les termes «Slavee» et «Slavic» ne diffèrent que par les deux dernières lettres. Par conséquent, si l'ensemble des suffixes définis comprend «ee» et «ic», il y aura perte d'information puisque les deux mots deviendront identiques. Bien qu'en règle générale, la suppression des suffixes donne des résultats satisfaisants dans la plupart des applications, ce n'est pas du tout le cas ici. On peut résoudre le problème en définissant les mots racine «Slave» et «Slavic»; en conséquence, la suppression des suffixes n'est pas exécutée dans ce cas précis. Ainsi, lorsqu'on découvre des cas où la suppression des suffixes pose un problème, il est possible de définir des mots racine et les substituts correspondants pour résoudre la difficulté.

Mots répétés: Enfin, le système examine la série de mots résultant de la stratégie d'analyse pour y découvrir les mots répétés.

Il convient de souligner que les mots identifiés comme répétés à ce stade peuvent ne pas avoir été reconnus comme tels avant l'exécution de la stratégie d'analyse. On ne conserve qu'une occurrence de chaque mot défini à ce stade.

Souignons que si l'apostrophe n'était pas supprimée, elle servirait très probablement de délimiteur. La première expression compterait alors trois mots et la seconde deux mots et les deux expressions n'auraient qu'un seul mot en commun.

Caractères de remplacement: Cette fonction est très utile pour uniformiser les abréviations. Une telle uniformisation est souhaitable car les abréviations renferment souvent des caractères qui, malgré leur utilité pour l'abréviation, passeraient pour des délimiteurs à une étape ultérieure de la stratégie d'analyse. Si tel était le cas, il y aurait presque inévitablement une perte d'information.

Par exemple, si l'on définissait pour la chaîne de caractères «T.V.» une valeur de remplacement «Télévision», la seconde expression serait substituée à la première chaque fois que celle-ci figurerait dans un texte.

Souignons que si l'on n'utilisait pas de caractères de remplacement, l'expression «T.V.» deviendrait très probablement, par la stratégie d'analyse, «T» et «V». Cela est tout à fait indésirable car l'abréviation perd ainsi toute sa signification.

Caractères alphanumériques: Dans le ACTR, un mot est défini comme une suite de caractères contenus dans la liste de caractères alphanumériques. Les caractères qui ne figurent pas dans cette liste servent de délimiteurs et ne sont pas pris en considération.

En règle générale, l'ensemble des caractères alphanumériques comprend toutes les lettres de l'alphabet et tous les caractères numériques. Compte tenu de cela, une expression comme «Farmer/Fisherman» donnera deux mots puisque «/» n'est pas un caractère alphanumérique et, de ce fait, sert de délimiteur.

Traitement de mots

À cette étape, le ACTR commence à traiter le texte comme une collection de mots. Par conséquent, les étapes qui suivent s'appliquent à chacun des mots pris individuellement.

Mots à trait d'union: Les mots à trait d'union sont remplacés par le ou les substitués prévus. Cette fonction est très utile en ce qu'elle permet de reconnaître des mots ou des groupes de mots auxquels la règle du trait d'union est appliquée de façon incohérente.

Si, par exemple, l'utilisateur définit «Take-Out» comme un mot à trait d'union avec comme substitut «Take-Out», la substitution s'effectuera. Par contre, si l'utilisateur ne prend pas soin de définir un substitut, «Take-Out» deviendra deux mots si le trait d'union n'est pas un caractère alphanumérique.

Caractères alphanumériques non valides: Si un mot est formé d'une chaîne de caractères qui le rendent intelligible, ce mot est supprimé sans autre considération.

Dans certaines applications, par exemple, on utilise cette fonction pour supprimer des mots qui renferment des caractères numériques. Par conséquent, si l'ensemble des caractères numériques étaient considérés comme des caractères non valides, un mot comme «DEPT716A» serait supprimé.

Mots de remplacement: Cette fonction prévoit l'utilisation de synonymes pour faire en sorte que deux mots dissimilables soient reconnus à des fins d'appariement. Cette fonction peut aussi être utile pour venir à bout des fautes d'orthographe courantes.

Par exemple, si l'on considérait les expressions «Automobile Repairs» et «Car Repairs» en sachant que le mot «Car» est le substitut de «Automobile», les deux expressions seraient classées identiques.

Mots doubles: Par cette fonction, le système ACTR considère non seulement l'occurrence d'un groupe de deux mots mais aussi l'ordre de ces mots. Cette fonction peut être utile pour résoudre les incohérences dans l'orthographe et conserver l'ordre des mots.

Considérons, par exemple, l'expression «Take Out Restaurant». Quoique cette expression donnerait trois mots parfaitement acceptables, les mots «Take» et «Out» ne concorderaient

présentant des différences syntaxiques et grammaticales devraient pouvoir devenir identiques grâce à la stratégie d'analyse. Si nous reprenons l'exemple ci-dessus, une stratégie d'analyse bien exécutée devrait transformer les deux expressions «COMPUTER PROGRAMMER» et «PROGRAMMING COMPUTERS» en deux expressions identiques. Par exemple, les deux expressions pourraient se ramener à «COMPUT PROGRAM».

La stratégie d'analyse peut comprendre la suppression des marques du pluriel, des mots sans importance et des suffixes et un certain nombre d'autres étapes. Bien que l'ordre des étapes soit déterminé par le ACTR, les utilisateurs contrôlent l'exécution de chaque étape. Pour plus de renseignements sur l'ordre des étapes de la stratégie d'analyse, le lecteur est prié de se référer à Connor, Salloum et Wenzowski (1988).

Essentiellement, la stratégie d'analyse comprend deux grandes composantes:

1. TRAITEMENT DE TEXTES. À ce stade, le texte en question est traité comme une suite ininterrompue de caractères. Bien que l'on reconnaisse qu'un texte renferme normalement des mots, des espaces et des signes de ponctuation, aucune attention particulière n'est accordée à ces éléments à ce stade-ci du processus. Cette approche est indispensable pour bien situer des chaînes particulières de caractères dans la phrase.
2. TRAITEMENT DE MOTS. Lorsque cette étape s'amorce, le texte est déjà divisé en mots et le traitement porte alors sur chacun des mots pris individuellement. Cette approche est nécessaire puisque la standardisation de textes s'opère en grande partie en fonction de mots définis.

Traitement de textes

Comme nous l'avons vu plus haut, cette étape de la stratégie d'analyse est exécutée sans égard au contexte. Les étapes ci-dessous sont donc exécutées en fonction des caractères pris individuellement.

Clauses d'exclusion: Les clauses d'exclusion ne sont pas appliquées dans l'appariement mais le sont dans la mise à jour de bases de données pour signaler l'intention de permettre la reproduction contrôlée d'expressions, à défaut de quoi le ACTR empêchera l'inclusion d'expressions identiques dans une base de données pour appariement.

En disposant d'un moyen pour contrôler la reproduction d'expressions, les utilisateurs peuvent inclure dans la base des expressions portant plusieurs codes même si ces expressions sont identiques après avoir fait l'objet d'une stratégie d'analyse. Bien qu'elles ne servent pas à l'appariement, les clauses d'exclusion sont incluses avec les expressions dans la base de données pour appariement et peuvent par la suite servir à résoudre manuellement des concordances multiples.

La syntaxe d'une clause d'exclusion est définie entièrement par l'utilisateur. Des caractères initiaux et terminaux doivent être définis. Durant l'appariement, on ne tient pas compte de ces caractères ni de l'information qu'ils renferment.

Considérons par exemple la syntaxe d'une clause d'exclusion définie par la suite de caractères initiaux «(Except» et le caractère terminal«)». Compte tenu de ces caractères, les deux expressions «Computer Programming (Except As An Employee)» et «Computer Programming (Except As Self-Employed)» pourraient se trouver simultanément dans la base de données pour appariement même si leurs formes standardisées sont identiques. Par conséquent, si l'on devait trouver une concordance pour «Computer Programmer», les deux expressions seraient signalées. Comme les clauses d'exclusion sont stockées avec les expressions originales dans la base de données, un analyste peut les visualiser et résoudre manuellement le problème de concordance.

Caractères d'annulation: Si un caractère d'annulation défini par l'utilisateur est repéré dans une expression, il en est éliminé par le ACTR en cours d'analyse.

C'est la façon dont on procède, par exemple, dans le traitement de texte anglais pour supprimer l'apostrophe. Ainsi, les deux expressions «Electrician's Apprentice» et «Apprentice Electrician» deviendraient identiques si l'on supprimait l'apostrophe.

2. UTILISATION DE ACTR

Pour utiliser le système ACTR, il faut tout d'abord définir les textes et les codes que l'on a l'intention d'utiliser comme données de référence pour l'appariement. Bien qu'il existe de nombreuses sources pour ce genre d'information, on choisira idéalement une série de textes qui sont représentatifs des textes ayant le plus de chances d'être observés dans une opération d'appariement. Dans le cas d'une enquête, il s'agira généralement des réponses fournies dans une enquête antérieure et des codes qui ont été attribués manuellement. Bien que l'on doive prendre soin de vérifier l'exactitude des codes qui ont été attribués, il est recommandé de conserver les textes dans leur forme intégrale, avec les erreurs d'orthographe, de grammaire et de syntaxe, puisque c'est sous cette forme que l'on risque de retrouver le plus souvent des textes dans des enquêtes ultérieures.

Après avoir défini un fichier de textes et de codes exacts, il faut intégrer ce fichier à une base de données servant à l'appariement. ACTR renferme le logiciel nécessaire pour exécuter cette tâche et convertit ainsi automatiquement le fichier en une base de données pour appariement.

ACTR a été conçu de manière à favoriser l'introduction de la technique d'itération dans les méthodes de codage automatique. Ainsi, il est possible d'ajouter, de modifier ou de supprimer des textes ou des codes à n'importe quel moment de l'application. De plus, on peut modifier la stratégie d'analyse (exposée en détail ci-dessous) à n'importe quel moment. Les utilisateurs ont donc à leur disposition un progiciel qui, par une succession d'opérations de mise à jour et d'appariement, permet d'effectuer autant d'itérations qu'il est nécessaire pour obtenir la qualité d'appariement voulue. Les utilisateurs sont invités à exploiter ACTR en ce sens puisqu'il se traduit finalement par une qualité supérieure et des opérations de codage plus économiques.

3. PRINCIPES DE FONCTIONNEMENT

Pour une personne qui exécute une opération de codage, les expressions «COMPUTER PROGRAMMER» et «PROGRAMMING COMPUTERS» (pour désigner une profession) sont tellement semblables qu'elles seront jugées identiques la plupart du temps. Bien que ce raisonnement semble tout à fait logique, un ordinateur ne sera pas porté à considérer ces deux expressions comme équivalentes. Malheureusement dans le langage humain (par exemple les langues anglaise ou française), il y a plusieurs façons d'exprimer la même chose. Par conséquent, si l'on veut qu'un ordinateur tienne compte de cette caractéristique du langage humain, il faut prévoir un mécanisme qui lui permettra de déterminer un degré de similitude entre des expressions.

Cela est l'essence même du système ACTR: une expression est évaluée en fonction de son degré de similitude avec une autre expression. Dans l'exemple ci-dessus, ACTR considère les deux expressions comme identiques puisqu'en supprimant les suffixes et les lettres doubles et en ne tenant pas compte de l'ordre des mots, on obtient dans les deux cas l'expression «COMPUT PROGRAM».

Les étapes qui ont abouti à l'expression «COMPUT PROGRAM» font partie de ce qu'on appelle dans le ACTR la stratégie d'analyse. Cette stratégie est entièrement sous le contrôle de l'utilisateur et peut être modifiée à n'importe quel moment d'une application. Les utilisateurs contrôlent la stratégie d'analyse en fournissant les données qui serviront à orienter le processus. Cela signifie que l'utilisateur exerce un contrôle parfait sur toutes les étapes de la stratégie, au point de décider de sauter une étape.

Stratégie d'analyse

La stratégie d'analyse est un processus du ACTR qui consiste à ramener des expressions à une forme standard. Idéalement, deux expressions comportant les mêmes mots mais

ACTR Un système généralisé de codage automatique

M.J. WENZOWSKI¹

RÉSUMÉ

Dans cet article, il est question de l'application généralisée d'une méthode de codage automatique. Jusqu'à récemment, le codage était une opération manuelle confiée à des personnes formées spécialement à cet effet; toutefois, la création de systèmes informatiques particuliers a contribué à éliminer sinon à réduire sensiblement le codage manuel. En règle générale, l'utilisation de ces nouveaux systèmes est limitée aux applications pour lesquelles ils ont été conçus. Le système qui est décrit ici peut servir à n'importe quelle forme de codage de textes anglais ou français selon n'importe quel mode de classification.

MOTS CLÉS: Codage automatique; classification; examen de textes.

1. INTRODUCTION

Le codage automatique est le procédé qui consiste à analyser un texte par ordinateur afin de lui attribuer un code. Pour être efficaces, les systèmes de codage automatique doivent pouvoir traiter les difficultés suivantes: nouvel ordre des mots, formes du singulier et du pluriel, absence de mots, présence de mots non pertinents, différences d'orthographe, synonymes, abréviations, emploi incohérent des traits d'union et application variée des règles de ponctuation et de syntaxe. De plus, lorsqu'il s'agit d'explorer une base de textes en vue d'un appariement possible en l'absence d'expressions identiques.

Les systèmes généralisés offrent toutes les fonctions voulues sous la forme d'un projeté souple et efficace. La personne qui souhaite se servir d'un système généralisé pour une application particulière n'a pas à élaborer ou à convertir quoi que ce soit pour adapter le système aux exigences particulières de l'application. En outre, il n'est pas nécessaire de financer la maintenance d'un système généralisé à même des frais aux utilisateurs puisque le projeté est à la charge d'un organisme central.

Le ACTR (acronyme pour Automated Coding by Text Recognition-codage automatique par reconnaissance de texte) utilise des techniques semblables à celles qui sont prévues dans d'autres systèmes de codage automatique en gestation à Statistique Canada (Landry et Pidcock 1984) mais il se distingue de tous les autres par le fait qu'il peut servir à n'importe quelle application visant à coder des textes anglais ou français selon n'importe quel mode de classification. Les méthodes utilisées par le système ACTR s'inspirent de méthodes qui ont été élaborées à l'origine au U.S. Bureau of the Census (Appel et Hellerman 1983). Essentiellement, la méthode consiste à examiner une série de textes préalablement codés. Si le texte en question est repéré, le code correspondant est enregistré et l'opération prend fin. Dans le cas contraire, l'examen se poursuit mais fait intervenir un algorithme pour repérer le texte le plus compatible; une fois cette opération réalisée, le système attribue le code correspondant.

¹ M.J. Wenzowski, Recherche et systèmes généraux, Statistique Canada, pièce 2306, Immeuble principal, Ottawa, Ontario, K1A 0T6.

BIBLIOGRAPHIE

- BOUCHON-MEUNIER, B. (1978). Sur la réalisation de questionnaires. Thèse d'Etat, Paris.
- KNAUS, R. (1987). Methods and problems in coding natural language survey data, *Journal of Official Statistics*, 3, 45-67.
- LORIGNY, J. (1982). Mesures d'entropie et d'information pour les systèmes ouverts complexes. Thèse d'Etat, Paris.
- LORIGNY, J. (1985). Manuel d'utilisation du système QUID. Institut National de la Statistique et des Etudes Economiques, Direction de la production, Paris.
- PICARD, C.-F. (1972). *Graphes et Questionnaires*. Paris: Gauthier-Villars.
- SHANNON, C.E. (1948). A Mathematical Theory of Communication. *Bell Systems Technical Journal*, 27, 379-423, 623-656.
- TERRENOIRE, M. (1970). Un modèle mathématique de processus d'interrogation: les pseudo-questionnaires. Thèse d'Etat, Grenoble.
- TOUNISSOUX, D. (1980). Processus séquentiels adaptés de reconnaissance de formes pour l'aide au diagnostic. Thèse d'Etat, Lyon.

Prenons un exemple (en partie fictif). Supposons que, selon la nomenclature, l'intitulé SECRÉTAIRE DE DIRECTION soit à chiffrer PCS = 4615 comme précédemment si la variable annexe CPF = 1 à 7, et PCS = 3726 (cadres de gestion courante des autres services administratifs des entreprises) si la CPF est égale à 8.

Considérons le FA contenant les deux références suivantes:

Intitulé				v.a. AE		v.a. CPF		code PCS	
Secrétaire	de direction	[49] 11		[83] 43		[7]		4615	
		[8]						3726	

Ces deux références sont donc bien correctement chiffrées. L'algorithme QUID arrivant à un sommet où il a tiré tout le parti possible des bigrammes de l'intitulé littéral, doit choisir maintenant un bigramme dans les variables annexes afin de séparer les deux issues finales PCS = 3726 et PCS = 4615. Dans notre exemple simple mais non dénué de réalisme, les trois bigrammes candidats AE1, AE2 et CPF apportent la même quantité d'information (un bit). La convention arbitraire prise dans notre algorithme est qu'en cas d'égalité, il choisit la première question dans l'ordre des déclarations des variables annexes. Ce qui dans notre exemple est trouvé un bon ordre des variables annexes qui éviterait ce défaut dans tous les cas de figure. On ne peut que chercher un ordre de déclaration statistiquement le moins mauvais (en tâtonnant à partir de l'ordre des clivages conceptuels, de la capacité néguentropique de chaque variable annexe, etc . . .)

5. CONCLUSION

Le système QUID dans sa version QUID 1 actuelle rend de précieux services à l'INSEE mais présente encore des points faibles dans le traitement des variables annexes.

La nouvelle version QUID 2 devrait améliorer ce traitement tout en restant fidèle à notre «approche de base» du problème de la codification automatique, que l'on peut résumer en deux points:

1. Séparation de la base d'apprentissage (ici, base de règles et de tables de décision écrites en clair, indépendantes les unes des autres, apurées et gérées par un atelier d'expertise autonome) et des logiciels de chiffrement automatique (ici, logiciels de chargement et d'exploration des tables).
2. Construction de logiciels généraux, c'est-à-dire indépendants du champ sémantique traité.

C'est du moins l'objectif que nous essaierons de maintenir.

REMERCIEMENTS

Je remercie les arbitres pour leur aide précieuse dans la rédaction de cet article.

Pour pallier ce défaut dans le système actuel, on ne peut qu'appliquer le contrôle de redondance aux variables annexes (et obtenir ainsi un cas douteux traité en rejet et correction manuelle au lieu d'un cas d'erreur passant inaperçu). Mais, là encore, ce n'est qu'un pis-aller. En effet, les variables annexes produisent un foisonnement anarchique dans le FA. Chaque référence du FA a sa propre combinaison croisée de modalités des variables annexes et il est peu probable de retrouver la même combinaison pour un nouvel individu à chiffrer. On aboutira donc à de nombreux cas d'échos douteux, donc à des rejets du chiffrement automatique, ce qui amoindrit le bénéfice pratique de l'exploitation de masse.

Les deux défauts n° 1 et n° 2 sont reliés à l'incomplétude relative du FA. Par exemple, il suffirait de placer dans le FA huit intitulés AGENT D'EXPLOITATION FORESTIÈRE complétés chacun par une des modalités possibles de CPF (1 à 8) pour que les deux défauts disparaissent. Mais hélas, dans les applications réelles, il se trouve que l'incomplétude relative du FA ne diminue que lentement au fur et à mesure de sa croissance jusqu'à un régime de croisière. Contrairement à l'espace lexicographique des intitulés littéraux qui, lui, tend à se densifier assez rapidement, l'espace croisé des variables annexes conserve très longtemps une vaste frontière passant lentement de la densité d'occupation 0 à la densité 1 (un individu).

Défaut n° 3. Il existe une troisième catégorie de difficultés qui tient non plus à l'incomplétude du FA mais à la sensibilité excessive de QUID par rapport aux erreurs inévitablement contenues dans le FA (et cela toujours pour ce qui concerne les variables annexes).

Prenons un exemple simplifié. Supposons que l'intitulé SECRÉTAIRE DE DIRECTION doive être chiffré PCS = 4615 (personnel de secrétariat de niveau supérieur) et ceci quelle que soit la valeur de toutes les variables annexes. Considérons le FA suivant dans lequel une erreur s'est glissée (par exemple une faute de saisie du code PCS):

Intitulé		v.a. CPF		v.a. AE		code PCS	
Secrétaire de direction	[7]	[49]	[1]	création de mode, haute couture)	(coopératives de crédit)	4616	erreur
Secrétaire de direction	[7]	[49]	[1]			4615	

Alors que la variable annexe AE ne devrait pas servir au chiffrement du code PCS, l'algorithme QUID va s'en emparer pour séparer les deux sommets de décision.

- L'un en faveur de 4615 au vu du bigramme AE1 = 49.
- L'autre en faveur de 4616 au vu du bigramme AE1 = 83.

Le résultat est qu'au stade du chiffrement proprement dit, toutes les secrétaires de direction appartenant à d'autres branches économiques que celles commençant par 49 ou 83 sortiront en «cas de réponse inconnue». En outre, celles de toutes les branches commençant par 83 produiront bien entendu des erreurs, mais c'est surtout le premier phénomène qui est gênant et «injuste» puisqu'il affecte un domaine bien plus large que celui de l'erreur initiale.

Défaut n° 4. Enfin, l'algorithme QUID actuel présente une rigidité excessive dans le choix de la question optimale. Le plus souvent, il en résulte une simple inversion de l'ordre des questions dans le cheminement de recherche, par rapport à l'ordre qu'aurait préféré le concepteur. L'effet est donc secondaire puisque le résultat final est identique. Mais il peut aussi se produire des distorsions plus graves.

- au premier étage, le QUID 1 réservé au traitement de l'intitulé littéral et produisant soit le code définitif (quand il est complètement déterminé par l'intitulé), soit un code interne désignant une règle ou une table de décision opérant sur les variables annexes pour achever le calcul.
- au second étage, les règles ou tables de décision achevant la détermination du code définitif.

Examen détaillé des difficultés rencontrées

Il se trouve que certaines nomenclatures particulièrement complexes comme le code PCS (Nomenclature des Professions et des Catégories socio-professionnelles) font appel à la combinaison d'un intitulé littéral et de plusieurs variables annexes.

Par exemple, le chiffrage du code PCS utilise la variable annexe Catégorie Professionnelle (en abrégé CPF). Voici la question telle qu'elle figurait dans le bulletin individuel du Recensement de la Population de 1982:

Indiquez la catégorie professionnelle de votre emploi actuel:

- manoeuvre ou manoeuvre spécialisée 1
- ouvrier spécialisé (OS, O1, O2, O3 . . .) 2
- ouvrier qualifié (P1, P2, P3, TA, OP, OQ . . .) 3
- employé 4
- technicien, dessinateur 5
- dirigeant des ouvriers ou des techniques 6
- dirigeant des agents de maîtrise ou des techniciens 7
- ingénieur ou cadre 8

L'adjonction de cette question subsidiaire est rendue nécessaire par le fait que l'intitulé seul ne suffit pas toujours à classer l'individu dans la nomenclature PCS.

- Par exemple un AGENT D'EXPLOITATION FORESTIERE
- doit être classé en 6916 (ouvriers d'exploitation forestière ou de sylviculture)
 - si sa CPF est 1, 2, 3 ou 4
 - et doit être classé en 4801 (Personnel de direction et d'encadrement des exploitations agricoles ou forestières)
 - si sa CPF est 5, 6, 7 ou 8.

Le système actuel considère ces variables annexes comme s'il s'agissait de données littérales. Elles sont placées à la fin de l'intitulé et structurées comme lui en bigrammes (par exemple, la variable CPF complétée par un blanc est placée dans le (m + 1)ème bigramme). Mais la solution n'est pas satisfaisante et plusieurs défauts apparaissent:

Défaut n° 1. L'insuffisance du FA conduit à de nombreux cas de réponse inconnue.

Par exemple, si le FA ne comprend qu'un AGENT D'EXPLOITATION FORESTIERE de CPF = 2 et un autre de CPF = 7 il ne pourra pas retrouver un AGENT D'EXPLOITATION FORESTIERE de CPF différent de 2 ou 7 (c'est-à-dire *a priori* dans 6 cas sur 8). Le défaut est aggravé lorsque la variable annexe est très diluée comme par exemple la variable Activité Economique de l'entreprise (en abrégé variable annexe AE).

Défaut n° 2. L'insuffisance du FA conduit à des cas d'erreur.

Par exemple, si le FA comprend un seul AGENT D'EXPLOITATION FORESTIERE, de CPF = 2, le bigramme CPF ne discrimine plus rien et ne figurera pas dans la clé de recherche, de sorte qu'un AGENT D'EXPLOITATION FORESTIERE de CPF = 7 sera classé en PCS = 6916 au lieu de 4801. C'est un cas d'erreur.

3.3 L'exploitation de chiffrement proprement dit

Pour chiffrer un intitulé de l'enquête en cours, on commence par le normaliser selon 3.1. Puis, les bigrammes obtenus sont apparés avec ceux du quid chargé dans l'ordinateur. L'exploration conduit à trois issues possibles.

3.3.1 Sommet de décision

Le système fournit un code unique mais qui peut être bien être erroné si la base d'apprentissage se trouve être trop pauvre. Par exemple, dans un de nos premiers essais en 1979, nous obtenions un sommet de décision de niveau 1, par bigramme 2 = CC, au vu du seul intitulé appris VACCINEUR VOLAILLES.

Lorsqu'un est apparu ensuite l'intitulé à chiffrer RACCOMMODEUR VÊTEMENTS, le code unique obtenu était celui des professions de service à l'agriculture et l'erreur était manifeste. Le système a donc été complété par une procédure de contrôle des échos uniques, dite «contrôle par la redondance» et consistant à vérifier après la détection d'un écho unique le contenu des trois premiers bigrammes de chaque mot. Un écho unique (issu du cheminement aboutissant à un sommet de décision) est déclaré non douteux quand il existe dans la grappe des intitulés du sommet de décision au moins un intitulé possédant les mêmes bigrammes de redondance que ceux de l'intitulé à chiffrer. Il est déclaré écho douteux dans le cas contraire et par conséquent traité comme anormalie du système automatique. L'expérience a montré que cet aménagement consolidait beaucoup la fiabilité du système sans alourdir notablement les tables en mémoire ni les temps de traitement (même dans les grosses applications, le nombre de formules de redondance par sommet de décision est en moyenne de l'ordre de l'unité et dépasse rarement la dizaine).

Pour être complet, ajoutons que ce contrôle par la redondance n'est pas figé une fois pour toutes. L'utilisateur dispose de deux paramètres externes: la liste des bigrammes sur lesquels il entend exercer le contrôle, et le nombre (maximum) des bigrammes retenus. Il peut ainsi doser la sévérité du contrôle d'appariement selon ses objectifs respectifs de qualité et d'«efficacité» du chiffrement automatique.

3.3.2 Un sommet d'indécision

Le système fournit plusieurs codes possibles (le plus souvent, deux codes) et affiche leurs fréquences d'occurrence respectives au sommet considéré. C'est un cas de rejet traité manuellement par l'agent disposant du dossier de l'enquête en cours de traitement.

3.3.3 Un cas de réponse inconnue

Lorsque, au cours de l'exploration du quid, la modalité recherchée ne se trouve pas dans les modalités apprises du bigramme interrogé, la recherche échoue. C'est aussi un cas de rejet à traiter manuellement.

Les cas nouveaux rencontrés au cours d'une exploitation sont mémorisés, puis centralisés dans l'atelier d'expertise, contrôlés, enfin incorporés au FA en vue d'une nouvelle version enrichie du quid.

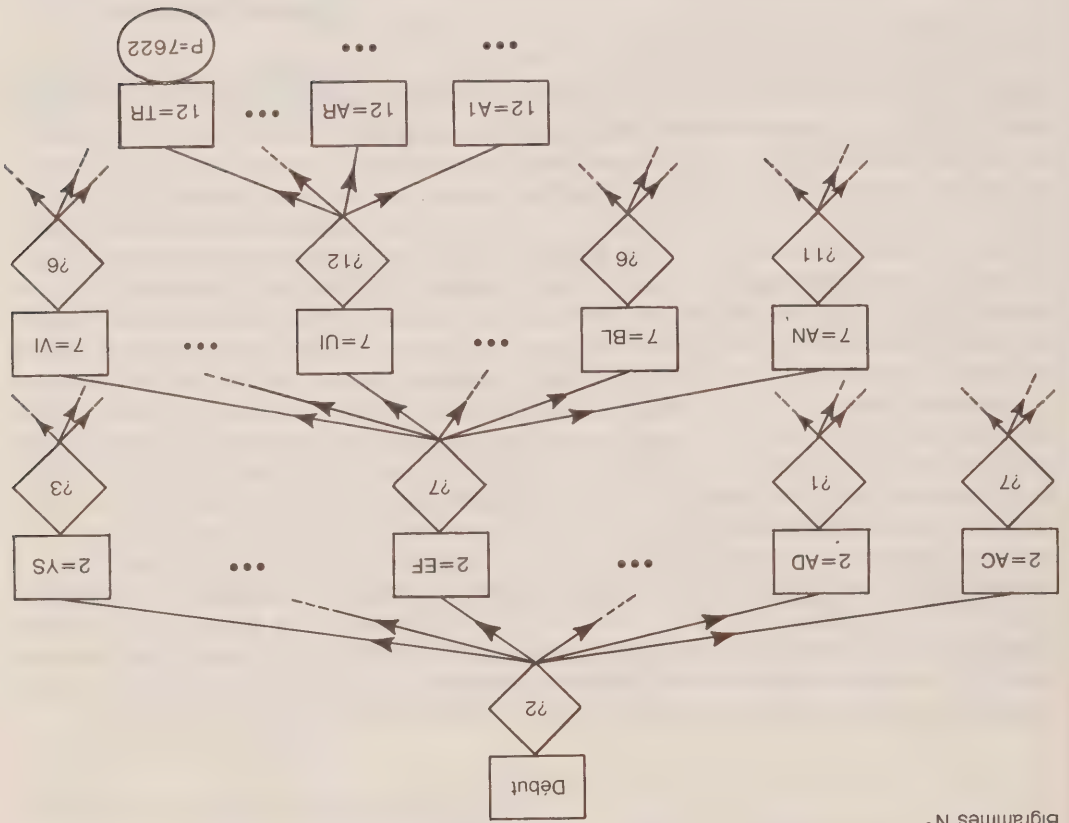
L'itération d'apprentissage se fait actuellement à un rythme annuel pour des raisons de commodité mais rien n'empêche de l'organiser à un rythme plus rapide pour une exploitation plus évolutive, telle que celle d'un Recensement de Population par exemple.

4. LE PROBLÈME DU TRAITEMENT DES VARIABLES ANNEXES

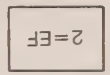
Dans la version actuelle QUID 1, les variables annexes sont simplement structurées en bigrammes et traitées comme des données littérales. Il en résulte des difficultés et des défauts qui nous conduisent à préparer une version QUID 2 fonctionnant à deux étages:

Intitulé normalisé:

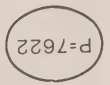
CH	EF					EQ	UI	PE				EN	TR	EI	IE	N
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		



En ce sommet de l'arborescence interroger le contenu du bigramme n° 2.



Le contenu du bigramme n° 2 est EF.



En ce sommet de l'arborescence, on peut déterminer le code Profession: sa valeur est 7622 (Nomenclature des métiers de 1975).

Dans l'exemple ci-dessus, l'intitulé brut est celui de la profession déclarée par l'individu recensé. L'objectif du système est de déterminer le Code Profession correspondant, dans la Nomenclature des Métiers de 1975.

Dans une première étape, on extrait les dix premiers caractères des trois mois les plus significatifs. On obtient ainsi l'intitulé normalisé qui est alors découpé en couples de lettres (appelés bigrammes) numérotés de 1 à 15. Ensuite commence l'interrogation proprement dite. Elle s'opère selon un enchaînement de questions-réponses optimisée par un algorithme mathématique basé sur la théorie de l'information. Ce calcul a lieu au cours d'une phase préliminaire qui détermine, en fonction du fichier d'apprentissage donné, le premier bigramme à interroger, puis la séquence des questions suivantes selon la réponse obtenue chaque fois. Ici l'ordinateur interroge d'abord le bigramme n° 2, qui contient EF, puis le bigramme n° 7, qui contient UI, et enfin le bigramme n° 12, qui contient TR. A ce stade il constate qu'il peut sans ambiguïté déterminer qu'il s'agit du code Profession 7622 (Agents techniques et techniciens s.a.). La durée totale du traitement est en moyenne de 41 millisecondes d'ordinateur IBM 370/148 et la mémoire centrale utilisée, de 380 K octets.

Intitulé brut: chef d'équipe d'entretien.

Figure 1. Exemple de classement d'un intitulé dans l'arborescence.

L'incertitude a priori sur T est mesurée par l'entropie de Shannon:

$$H(T | x_0) = \sum_j \Pr(t_j | x_0) \log 1/\Pr(t_j | x_0).$$

Supposons qu'un bigramme (quelconque) q_i soit affecté au sommet x_0 . À chacune des modalités qu'il prend dans le FA nous associons la sous-base constituée des intitulés possédant cette modalité.

Notons $(a_1^2, a_2^2, \dots, a_k^2, \dots)$ les modalités prises par le bigramme q_i dans le FA. Pour chacune de ces modalités, donc pour chacune des sous-bases engendrées, nous créerons un sommet y , successeur immédiat de x et placé au niveau 1 de l'arborescence.

L'information apportée par le bigramme q_i (supposé affecté au sommet-racine x_0) est mesurée par la réduction moyenne de l'incertitude sur T en passant de x_0 à l'un des sommets y .

Soit:

$$I(x_0, T, q_i) = H(T | x_0) - \sum_{y \in T(x_0)} \Pr(y) H(T | y),$$

où l'on note

$T(x_0)$ l'ensemble des sommets y successeurs au niveau 1 du sommet x_0

$H(T | y)$ l'entropie conditionnelle de T au sommet y .

(même formule que ci-dessus en remplaçant x_0 par y).

$\Pr(y) = N(x_0, a_i^2) / N(x_0)$ si a_i^2 est la modalité du bigramme q_i qui engendre le sommet y et $N(x_0, a_i^2)$ la fréquence d'occurrence de la modalité a_i^2 du bigramme q_i dans le FA

tout entier.

L'algorithme effectue ce calcul d'information pour tous les bigrammes q_1, q_2, \dots, q_m , puisqu'au sommet racine x_0 ils sont tous candidats possibles à la sélection comme premier bigramme interrogé.

L'algorithme choisit le bigramme qui maximise $I(x_0, T, q_i)$, notons le q_{i_0} , et partage alors effectivement la base en autant de sous-bases que de modalités du bigramme q_{i_0} rencontrées dans la base. Les sommets y , successeurs de x_0 au niveau 1 sont alors effectivement créés. La construction du niveau 1 de X est terminée.

Pour chaque sous-base obtenue (donc pour chaque sommet y) l'algorithme recommence exactement le même traitement que celui que nous venons de décrire à propos du sommet-racine x_0 etc. etc.

Le processus s'arrête pour un sommet déterminé:

- (1) lorsqu'il n'y a plus qu'un intitulé au sommet, et dans ce cas l'entropie conditionnelle équivaut à zéro, ou
- (2) lorsqu'il n'existe qu'un nombre restreint d'intitulés qui diffèrent en ce qui a trait aux bigrammes restants, mais qui possèdent tous le même code, ou
- (3) lorsqu'il existe deux intitulés ou plus mais qui possèdent des codes différents et non distinguables.

Les cas (1) et (2) sont dits «sommets de décision», le cas (3), «sommet d'indécision». Ils constituent ensemble les «sommets terminaux».

La progression de la construction de l'arborescence X se poursuit de niveau en niveau jusqu'à épuisement du FA. En fait, nous n'avons jamais dépassé le niveau 15 mais aucune limite n'est fixée par le système lui-même. Un exemple de classement dans l'arborescence est donné en

figure 1.

3. LE SYSTÈME QUID DANS SA VERSION ACTUELLE (ou QUID 1)

3.1 Normalisation préalable des intitulés

Avant de construire l'arborescence optimisée, les intitulés bruts subissent d'abord un traitement automatique de normalisation préalable, commandé par un jeu de paramètres externes choisis par l'utilisateur pour son application.

Les mots sont séparés et cadrés dans des zones fixes dont la longueur (unique pour tous les mots) et le nombre maximum (unique pour tous les intitulés) sont paramétrés. Il est conseillé de choisir par ces deux paramètres une valeur plutôt large et de laisser l'algorithme d'optimisation sélectionner lui-même les parties significatives de l'intitulé (cf. 3.2). Par exemple, l'application des DADS (cf. 2.2) a choisi 4 zones de 12 caractères chacune.

Les «mots vides» sont éliminés. La liste des mots vides est un paramètre externe fourni par l'utilisateur pour son application. Elle comprend le plus souvent les articles, prépositions, *etc.* et dépend beaucoup de l'application.

Les sigles sont normalisés (I.N.S.E.E. devient INSEE, S.N.C.F. devient SNCF). Enfin l'utilisateur peut effectuer sur la table des mots séparés tout traitement particulier de son choix (sous forme d'un sous-programme en langage PL/1). En fait, cette possibilité apparaît rarement nécessaire et est très peu utilisée (sauf pour le chiffrement des codes de commune à partir des intitulés de commune).

Lorsque le traitement des mots est terminé, ceux-ci sont découpés en bigrammes (tranches de deux lettres consécutives) ou trigrammes (tranches de trois lettres consécutives), ou *etc.* Le choix de ce mode de découpage est paramétré (mais unique pour toute l'application traitée). En pratique, le découpage en bigrammes est le seul à avoir été utilisé jusqu'à présent mais l'idée d'un découpage en trigrammes mériterait d'être expérimentée. Pour la suite de l'exposé, nous considérerons uniquement le découpage en bigrammes.

3.2 L'algorithme de construction de l'arborescence optimisée

Notons $T = (t_1, t_2, \dots, t_j, \dots, t_n)$ le code à chiffrer, par exemple l'ensemble des modalités du code Profession.

$\tilde{Q} = (q_1, q_2, \dots, q_i, \dots, q_m)$ l'ensemble des bigrammes résultant de la normalisation des intitulés (par exemple $m = 24$ si l'on a choisi le nombre 4 comme paramètre «nombre de mots» et 12 caractères comme paramètre «longueur de mot»).

$X =$ l'arborescence à construire, que nous appelons un «quid» (questionnaire d'identification).

L'algorithme construit X en descendant du sommet-racine x_0 (placé par convention au «niveau 0») jusqu'aux sommets de niveaux 1, 2, *etc.*

Au sommet-racine x_0 il associe le FA tout entier et cherche le meilleur bigramme à intercaler en premier, c'est-à-dire celui qui, dans le FA tout entier est le plus discriminant pour le code cherché T .

Notons $N(x_0)$ la fréquence d'occurrence totale associée au FA entier, c'est-à-dire la somme des fréquences accompagnant les intitulés de la base,

$N(x_0, j)$ la fréquence d'occurrence du code t_j dans le FA tout entier.

Nous supposons la population d'apprentissage statistiquement représentative de la population à chiffrer (rappelons que le FA est très souvent, en pratique, le fichier d'enquête d'une année antérieure).

On peut donc estimer la probabilité de trouver le code t_j dans la population à chiffrer, par la formule:

$$\Pr(t_j | x_0) = N(x_0, j) / N(x_0).$$

attribué par un expert. La base de données est la plus étendue possible en vue de permettre l'atteinte d'un taux élevé d'appariement, et l'on ajoute à la base de nouveaux intitulés à mesure que ceux-ci apparaissent.

Dans notre terminologie, la base de données s'appelle «base d'apprentissage», ou «fichier d'apprentissage» (FA) parce qu'elle présente à l'état brut la structure ordinaire d'un fichier plat. Pour constituer le fichier d'apprentissage, nous partons le plus souvent de l'enquête d'une année antérieure, déjà chiffrée manuellement ou par une méthode interactive. Chaque intitulé de la base est accompagné de son code (supposé *a priori* exact), et de sa «fréquence d'occurrence» observée dans le FA, c'est-à-dire du nombre d'individus ayant répondu par cet intitulé. La tâche de gestion de la base d'apprentissage (apurement, extension) est complètement déconnectée de l'exploitation de chiffrement de l'enquête en cours. Elle est confiée à un atelier centralisé composé de codeurs spécialisés, tandis que l'exploitation de chiffrement proprement dite est le plus souvent décentralisée régionalement.

La difficulté propre à une approche de ce type provient de l'accroissement rapide du temps de recherche dans la base au fur et à mesure que sa taille augmente. Pour y remédier, le système QUID utilise des résultats mathématiques de la Théorie de l'Information (Shannon 1948; Picard 1972; Bouchon-Meunier 1978; M. Terrenoire 1970; Tounissoux 1980), grâce auxquels le temps de recherche est minimisé en organisant la base sous forme d'une structure arborescente optimisée.

L'approche de base du système QUID permet aussi d'opter pour un ensemble de logiciels généraux, c'est-à-dire s'appliquant à tous les champs sémantiques, comme par exemple des intitulés de profession, des intitulés de produits alimentaires, ou des intitulés de communes.

2.2 Les résultats obtenus

Le système est expérimenté sur différents travaux de l'INSEE et fonctionne en exploitation courante pour le chiffrement du code CS (catégorie socio-professionnelle) dans le traitement des DADS (Déclarations annuelles de données sociales) fournies par toutes les entreprises employant des salariés. Indiquons quelques chiffres pour situer les ordres de grandeur.

Dans l'application aux DADS, le fichier d'apprentissage comprend à ce jour 122 000 intitulés (représentant une population d'apprentissage de 650 000 salariés). Son organisation optimisée est une arborescence d'environ 100 000 sommets (dont 86 000 sommets de décision cf. 3.2). Il a été utilisé pour chiffrer une population de 570 000 salariés avec une efficacité moyenne de 90%, variant entre 85% et 95% selon les régions. Nous entendons par «efficacité» le pourcentage de cas où le système fournit une réponse unique, que nous acceptons par principe dans les conditions de cette application. Ne disposant pas actuellement de mesure précise de la validité de ces réponses uniques, nous estimons vraisemblable un taux d'erreur de l'ordre de 5% à 10%. Toutefois, la base d'apprentissage est en cours d'apurement à l'atelier d'expertise de Dijon, après quoi le taux d'erreur devrait normalement diminuer dans une proportion notable. Nous aurons des chiffres plus précis à communiquer à ce moment-là.

Du point de vue des contraintes informatiques, l'arborescence optimisée est chargée dans 3 300 kilooctets de mémoire centrale (virtuelle) et le temps de chiffrement automatique d'un cas individuel est de l'ordre de 40 ms d'unité centrale IBM 4341.

Nous possédons depuis quelques mois une variante du logiciel de chiffrement proprement dit destinée aux mini-ordinateurs et qui charge l'arborescence par parties, en fonction de l'espace mémoire autorisé.

Dans d'autres applications que celle des DADS, l'efficacité est moindre et ne dépasse pas 75%. Tout dépend de la qualité et de l'exhaustivité de la base d'apprentissage.

QUID, une méthode générale de chiffrement automatique

JACQUES LORIGNY¹

RÉSUMÉ

Le système QUID, conçu et développé par l'INSEE (Paris) est un système de chiffrement automatique de données d'enquête recueillies sous forme d'initiales littérales exprimées dans la terminologie du répondant. Le système repose sur l'utilisation d'une très vaste base d'apprentissage composée de phrases réelles codifiées par des experts. L'article présente d'abord le traitement automatique de normalisation préalable des phrases, puis l'algorithme organisant la base de phrases en une arborescence optimisée. Un exemple de classement est donné en illustration. Le traitement des variables annexes de codification, venant compléter l'information contenue dans les phrases, présente actuellement des difficultés qui sont examinées en détail. Le projet QUID 2, version renouée du système, est évoqué succinctement.

MOTS CLÉS: Codification automatique; variables en langue naturelle; appartenance de phrases; N-grammes.

1. INTRODUCTION

Le système QUID (abrégié de Questionnaires d'Identification) est un système de chiffrement automatique conçu et développé par l'Institut National de la Statistique et des Études Économiques (INSEE) depuis les années 1979-1980.

Rappel du problème

Le problème consiste à classer automatiquement un individu enquêté dans un poste défini d'une nomenclature existante (par exemple, la nomenclature des Professions). Pour cela, le système utilise principalement la réponse en clair à la question posée directement (par exemple «Quelle profession ou quel métier exercez-vous actuellement ?»), et accessoirement d'autres informations figurant dans le formulaire d'enquête et supposées préalablement codifiées (par exemple, le code Activité Économique de l'entreprise employant l'individu).

Dans notre terminologie, la réponse directe en clair est appelée «initiale littérale», ou en abrégé «initiale». Les informations codifiées complémentaires sont désignées sous le terme générique de «variables annexes». Nous présentons dans la prochaine section l'approche de base du système QUID, et donnons des résultats de son application à l'INSEE. Dans la section 3, nous décrivons le système dans sa version actuelle. Enfin, nous examinons le problème du traitement des variables annexes dans la section 4. La nouvelle version, QUID 2, présentée dans cette même section, devrait aider au traitement des difficultés rencontrées.

2. LE PRINCIPE DE LA MÉTHODE

2.1 L'approche de base

L'approche de base du système QUID consiste à élaborer une base de données très importante constituée d'initiales typiques des répondants, accompagnées du code correspondant

¹ Jacques Lorigny, Administrateur à l'Institut National de la Statistique et des Études Économiques 18, Bd Adolphe Pinard 75675 PARIS CEDEX 14 (France).

- COX, L. (1987). A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association*, 82, 520-524.
- COX, L., FAGAN, J., GREENBURG, B., et HEMMIG, R. (1986). Research at the Census Bureau into disclosure avoidance techniques for tabular data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 388-393.
- CRESSIE, N., et READ, T.R.C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society, Série B*, 46, 440-464.
- DEMING, W.E., et STEPHAN, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11, 427-444.
- DIFFENDAL, G. (1988). Test des opérations de redressement de 1986 dans le Central Los Angeles County. *Techniques d'enquête*, 14, 75-92.
- FAY, R.E. (1986). Implications of the 1980 PEP for future census coverage evaluation. U.S. Bureau of the Census, non publié.
- FAGAN, J.T., et GREENBERG, B. (1988). Algorithms for making tables additive: Raking, Maximum Likelihood, and Minimum Chi-square. *Proceedings of the Section on Survey Research Methods, American Statistical Association* (à paraître).
- GASS, S.I. (1964). *Linear Programming: Methods and Applications*. New York: McGraw-Hill.
- IRELAND, C.T., et KULLBACK, S. (1968). Contingency tables with given marginals. *Biometrika*, 55, 179-188.
- LITTLE, R.A., et RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.
- NATIONAL ACADEMY OF SCIENCES (1985). *The Bicentennial Census: New Directions for Methodology in 1990*. Washington: National Academy Press.
- OH, H.T., et SCHEUREN, F.J. (1978). Multivariate ratio raking estimation in the 1973 Exact Match Study. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 716-722.

- PURCELL, N.J. (1979). Efficient estimation for small domains: a categorical data analysis approach. Thèse de doctorat, University of Michigan.
- PURCELL, N.J., et KISH, L. (1979). Estimation for small domains. *Biometrics*, 35:365-384.
- RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.
- SCHEUREN, F.J. (1981). Methods of estimation for the 1973 exact match study. Dans *Studies from Interagency Data Linkages*, Washington: Social Security Administration.
- ZASLAVSKY, A.M. (1989). Representing Census undercount at the household level. Thèse de doctorat, Département de mathématiques, Massachusetts Institute of Technology, Cambridge, Massachusetts.

Nous devons avoir $Ha_{j,min} \leq b_j \leq Ha_{j,max}$ où $a_{j,min}$ et $a_{j,max}$ sont respectivement les valeurs minimum et maximum de a_j . Si tel n'était pas le cas, aucun poids ne pourrait satisfaire la contrainte f . Il doit donc y avoir au moins une racine ρ , et si les a_j sont non négatifs, l'expression s'accroît en ρ et cette racine est par conséquent unique. On détermine alors la valeur réelle de ρ par la méthode de Newton ou par une expression en forme analytique pour les racines d'un polynôme (puisque dans le cas de la matrice A , a_j désigne le nombre de membres de la classe j dans un ménage, nombre qui dépasse rarement 2).

Bien que nous n'ayons pas encore démontré que cet algorithme converge constamment, nous avons constaté son efficacité dans la pratique. Cet algorithme n'exige pas d'inversion de matrice et si les a_j sont de petits nombres entiers, le nouveau calcul des poids à chaque étape n'exige que le calcul de quelques puissances en nombre entier. De plus, si certaines contraintes ont la forme de fréquences marginales simples, le redressement dans ce cas se fait par l'itération classique.

Si la matrice A originale est utilisée, la méthode peut tirer parti de la faible densité de A (attribuable au fait que seulement quelques classes sont représentées dans chaque ménage). À chaque étape (par exemple, redressement en vue de l'ajustement à la fréquence marginale b_j), seuls les poids correspondant aux valeurs de a_j non nulles doivent être modifiés; par conséquent, chaque cycle n'exige que S_1 multiplications (le nombre d'éléments non nuls de A , qui est moindre que la population de l'ilot) et peut-être $3S_1$ additions, comparativement à $S_1 + S_2$ opérations par itération pour ce qui est de la méthode de Newton. Par ailleurs, comme les lignes de A tendent à être fortement dépendantes, la convergence peut être lente (en règle générale, 20 cycles dans nos simulations); l'orthogonalisation de A a pour effet de détruire la structure à faible densité des coefficients. Ainsi, à moins que S_2 soit beaucoup plus grand que S_1 (ou que l'on trouve une autre méthode pour activer l'algorithme), la méthode itérative du quotient n'est pas plus rapide que la méthode de Newton.

REMERCIEMENTS

Cette étude a été rendue possible grâce aux conventions 86-8 et 87-7 (Joint Statistical Agreements) entre le U.S. Bureau of the Census et l'Université Harvard sur la statistique. L'auteur a mis à profit les commentaires reçus de Donald B. Rubin et des autres participants au colloque sur le sous-dénombrement dans le recensement, organisé par le département de statistique de l'Université Harvard, de même que les commentaires reçus de Nathaniel Schencker, de Nash Monsoor et d'autres membres du Undercount Research Staff au U.S. Bureau of the Census. Les données ayant servi aux simulations ont été fournies par le Consortium interuniversitaire pour la recherche politique et sociale et avaient été recueillies à l'origine par le Bureau of the Census, Département du commerce des Etats-Unis.

BIBLIOGRAPHIE

ALEXANDER, C.H. (1987). Une classe de méthodes utilisant des chiffres de population dans la pondération des ménages. *Techniques d'enquête* 13, 193-209.

BISHOP, Y.M.M., FIENBERG, S.E., et HOLLAND, P.W. (1974). *Discrete Multivariate Analysis*. Cambridge: M.I.T. Press.

BUREAU OF THE CENSUS (1985). Census of Population and Housing, 1980: Public Use Microdata Samples.

CILKE, J.M., et WYSCARVER, R.A. (1988). The Individual Income Tax Simulation Model, dans *Compendium of Tax Research* 1987, Office of Tax Analysis. Washington: Government Printing Office.

Tableau 9

Comparaison de fonctions objectif pour l'ajustement de tableaux

Nom de la méthode d'ajustement	Fonction objectif $T_f(W)$, forme usuelle	Fonction objectif $T_0(W)$, uniformisée	Dérivée seconde, $T''_0(W)$
Moindres carrés (variance minimum)	$(W - 1)^2$	$(W - 1)^2/2$	1
Méthode itérative du quotient	$W \log W$	$(W \log W) - W + 1$	$1/W$
Maximum de vraisemblance	$-\log W$	$W - 1 - \log W$	$1/W^2$
χ^2 minimum	$(W - 1)^2/W$	$(W - 1)^2/2W$	$1/W^3$

vraisemblance ou à celle du χ^2 minimum, c'est surtout à cause de la forme élémentaire de sa solution et de l'analogie avec l'itération dans les tableaux de contingence. Cressie et Read (1984) étudient systématiquement les propriétés de cette catégorie de mesures d'ajustement.

10.2 Méthode de la descente cyclique pour l'ajustement de poids

Dans cette section, nous présentons une méthode d'ajustement analogue à l'ajustement proportionnel itératif (API) appliqué dans les tableaux de contingence. Selon l'API, les fréquences de case sont transformées par multiplication de manière à ne pas modifier les produits croisés (condition nécessaire pour minimiser la fonction objectif) et à rendre le tableau compatible avec chaque série de contraintes marginales. L'algorithme converge vers un tableau qui satisfait toutes les contraintes, et préserve nécessairement les produits croisés. (Bishop, Feinberg et Holland 1974; Ireland et Kuiliack 1968).

Selon la méthode proposée, les poids doivent avoir la forme linéaire logarithmique $W_i = \exp(a_i/\lambda - 1)$ définie dans la Section 3.3, et satisfaire en même temps les contraintes $AW = B$. Dans cet exposé, nous supposons que la contrainte de poids total $\sum W_i = H$ est exclue de $AW = B$ et que A est de dimension p (contraintes) $\times I$ (nombre de modes de composition des ménages). Nous allons exécuter une série d'étapes dans lesquelles nous multiplierons à chaque fois chacun des poids W_i par cp_{ajl} pour obtenir un nouveau poids W'_i , ce qui aura pour conséquence de conserver la structure linéaire logarithmique; c et p sont choisis de manière que les contraintes $\sum W'_i = H$ et $\sum W'_i/a_{ij} = b_j$ soient satisfaites. Si nous procédons de façon cyclique, de manière que $j = 1, 2, \dots, p$ désigne les contraintes l'une après l'autre, l'algorithme convergera éventuellement vers des poids qui satisfont toutes les contraintes.

À l'étape j du cycle l , les nouveaux poids sont définis $W'^i_{(l,j)} = cp_{ajl}W^i_{(l,j-1)}$ (initialisés pour $j = 1$ par l'utilisation des derniers poids du dernier cycle, $W^i_{(l,0)} = W^i_{(l-1,p)}$). Par conséquent c et p doivent satisfaire les équations suivantes:

$$\sum_{i=1}^I cp_{ajl} W^i_{(l,j-1)} = H, \quad \sum_{i=1}^I a_{ij} cp_{ajl} W^i_{(l,j-1)} = b_j. \tag{5}$$

Si nous supprimons c de ces équations, p est une racine de

$$\sum_{i=1}^I \left(H a_{ij} - b_j \right) W^i_{(l,j-1)} p^{a_{ij}} = 0. \tag{6}$$

les études révélèrent que le «biais dû à la composition» représente un problème appréciable, il faudrait élaborer des paramètres statistiques plus raffinés (Section 8.1). On devrait demander à un échantillon de ménages de l'EP qui ont été oubliés dans le recensement de répondre au questionnaire complet de manière à pouvoir mettre en modèle pour le redressement (Sections 8.2, 8.3) le rapport entre le sous-dénombrement et des covariables comme le revenu et le niveau d'instruction.

9.2 Applicabilité des redressements

On devrait tester et comparer les méthodes de la Section 5 à l'aide de données de l'EP.

9.3 Imputation multiple

Bien que les méthodes proposées dans cet article soient de nature déterministe, les statistiques fondées sur des enregistrements pondérés renferment toujours un certain nombre de sources d'incertitude, comme l'incertitude dans l'estimation des taux de sous-dénombrement, la variabilité des taux de sous-dénombrement de classes d'un flot à l'autre par rapport à la moyenne nationale, la variabilité binomiale du nombre réel de personnes ou de ménages oubliés par rapport au nombre prévu, étant donné le taux de sous-dénombrement, et l'incertitude concernant les écarts entre les valeurs de covariables pour les ménages oubliés et celles pour les ménages recensés qui sont pondérés à la hausse en vue de remplacer les premiers.

À des fins de recherche, on pourrait préparer des fichiers qui représenteraient toutes ces formes d'incertitude par l'imputation multiple (Rubin 1987). On pourrait représenter deux versions (ou plus) de la série de données repondérées en incluant plusieurs séries de poids dans le fichier. Les chercheurs pourraient reprendre leurs analyses en se servant à chaque fois d'une nouvelle série de poids. La variabilité des résultats des diverses versions donnerait une estimation de la variabilité attribuable au redressement pour sous-dénombrement. Zaslavsky (1989) analyse des méthodes d'imputation multiple dans ce contexte.

10. MÉTHODES ADDITIONNELLES

10.1 Choix d'une fonction objectif pour la pondération

On a proposé un certain nombre de fonctions objectif pour déterminer un tableau de données ajustées optimales (il s'agit le plus souvent de tableaux de contingence, voir Fagan et Greenberg 1988). Dans chaque cas, la fonction est de la forme $T = \sum T_i(W_i)$, où T_i prend une des formes indiquées dans le tableau 9. On peut ramener chacune de ces fonctions à une fonction équivalente T_0 en multipliant par un coefficient constant et en ajoutant une fonction linéaire de W de sorte que $T_0(1) = 0$, $T_0'(1) = 0$. Comme $\sum W_i$ est limitée à une valeur donnée, les poids optimaux ne sont pas modifiés. Les deux premiers termes des fonctions objectif normalisées développées en une série de Taylor de centre 1 sont les mêmes dans chaque cas et ces fonctions devraient produire des résultats comparables lorsque les poids se rapprocheront de 1. Par contre, elles diffèrent l'une de l'autre pour ce qui a trait au degré d'asymétrie entre les coûts de la pondération à la baisse et ceux de la pondération à la hausse, ce degré d'asymétrie étant défini par l'exposant de W dans la formule de dérivée seconde, $T_0''(W) = W^{-k}$. La méthode des moindres carrés $k = 0$ traite la pondération à la hausse et à la baisse de façon parfaitement symétrique et, de ce fait, peut produire des poids nuls ou négatifs. À mesure que k augmente, le coût de la pondération à la hausse diminue par rapport à celui de la pondération à la baisse. Toutes les autres fonctions objectif ($k > 0$) produisent un poids positif pour chaque élément des données brutes; cela est évident dans le cas de la méthode itérative du quotient, si l'on en juge par la forme des poids illustrée dans la Section 3.3. Si l'on préfère ici la méthode itérative du quotient à la méthode du maximum de

personne était déclarée. Le revenu de cette personne est susceptible d'être moins élevé que celui d'un membre adulte permanent d'une famille qui ne dépend pas de l'aide sociale. Par conséquent, ni le revenu du ménage recensé ni celui du ménage «rajouté» ne constituent une valeur satisfaisante pour l'imputation du revenu du ménage redressé.

Comme aucune correspondance directe n'est établie entre les ménages qui sont pondérés à la baisse et ceux qui reçoivent un poids additionnel, il n'est pas possible de reporter directement sur un ménage «rajouté» le revenu non redressé d'un ménage recensé. Toutefois, si l'on entreprenait des études visant à comparer les revenus des ménages recensés et des ménages oubliés, on pourrait se servir des revenus des ménages pondérés à la baisse pour le redressement des revenus. Par exemple, on pourrait poser la condition que le revenu moyen des ménages de l'ilot repondéré égale le revenu moyen des ménages de l'ilot avant redressement.

8.3 Redressement par pondération des caractéristiques non classificatoires

Supposons que nous disposions de données sommaires redressées (par ilot) sur quelques caractéristiques des ménages, exception faite des effectifs des classes de redressement. Par exemple, un modèle de régression pourrait nous avoir fourni une estimation redressée du revenu moyen ou de la proportion des familles monoparentales. Dans la mesure où l'on peut représenter les données sommaires comme une somme pondérée des valeurs de covariables pour chaque ménage, il est possible d'obtenir la conformité avec la valeur redressée voulue en définissant une liaison linéaire pour les poids, qui peut être intégrée à la méthode de redressement par pondération étudiée dans cet article. Par conséquent, si nous représentons l'exemple du revenu, nous définirions la contrainte de manière que la somme pondérée des revenus égale le produit du nombre de ménages par le revenu moyen redressé. Pour redresser la proportion de familles monoparentales, nous définirions la contrainte de manière que la somme pondérée des revenus égale le produit des indicateurs 0-1 pour cette caractéristique égale le nombre total voulu.

8.4 Résumé et incidences

La méthode proposée permettra de pondérer à la hausse des ménages et de fonder les analyses uniquement sur les caractéristiques de ces ménages sans que l'on s'interroge sur la possibilité d'un biais. Si le redressement et les biais introduits dans les caractéristiques des ménages sont de faible importance, le biais global contenu dans les caractéristiques estimées de l'ilot sera de second ordre et ne devrait pas poser de problème majeur. On pourrait peut-être réduire davantage les biais en effectuant quelques ajustements par régression simple.

9. PROPOSITIONS POUR ORIENTER LA RECHERCHE ET PERFECTIONNER LA METHODE

Cette section résume un certain nombre de propositions concernant l'application et le perfectionnement de la méthode de redressement étudiée dans cet article.

9.1 Collecte de données et modélisation statistique dans le cadre d'une enquête postcensitaire (EP)

L'EP devrait distinguer le sous-dénombrement des personnes dans les ménages recensés du sous-dénombrement des personnes dans les ménages oubliés et les deux taux de sous-dénombrement devraient faire l'objet de modèles distincts pour chaque classe de redressement. Les taux de sous-dénombrement des ménages devraient aussi faire l'objet de modèles. (Section 4). Une série de mesures (comme dans la Section 6.2) pourraient servir à comparer la composition des ménages «rajoutés» et celle des ménages oubliés qui ont été relevés dans l'EP; si

Par conséquent, si les ménages ayant un mode de composition donné sont sous-dénombrés dans une plus forte proportion que d'autres, ils peuvent être sous-représentés dans les listes pondérées et s'ils sont associés, par exemple, à un faible niveau de revenu, les estimations du revenu total seront biaisées par excès.

Le problème consiste essentiellement en un manque d'ajustement probable du modèle aux tendances des données. Les biais les plus prononcés pourraient se trouver dans les statistiques qui ont trait directement à la composition des ménages, comme le nombre de familles monoparentales.

Si le biais dû à la composition des ménages devait poser un problème sérieux, on pourrait tenter de le réduire en augmentant les taux de redressement des classes à l'aide de données additionnelles sur le taux de sous-dénombrement combiné des personnes de diverses classes (ou de classes groupées).

8.2 Biais de réponse

Étant donné un groupe de ménages ayant la même composition, il n'est pas déraisonnable de croire que ceux qui ont été oubliés dans le recensement diffèrent systématiquement, à certains points de vue, de ceux qui ont été recensés. Autrement dit, le sous-dénombrement peut être une forme de non-réponse avec biais. Par exemple, les ménages caractérisés par un faible revenu et un niveau d'instruction inférieur peuvent être plus susceptibles d'être oubliés entièrement ou partiellement; le revenu et le niveau d'instruction n'étant pas des variables classificatoires, elles ne font donc pas l'objet d'un redressement direct.

Selon les méthodes proposées, le *redressement de ménages entiers* s'effectue par une pondération à la hausse des ménages, sans modification des valeurs des covariables. On suppose implicitement que les ménages oubliés ne diffèrent pas, au point de vue de ces covariables, des ménages recensés qui ont la même composition. Aucune des données relatives à l'ilot faisant l'objet du redressement ne vient réfuter cette hypothèse. Toutefois, l'EP (enquête postcensitaire) devrait pouvoir nous renseigner sur les différences entre les ménages recensés et les ménages oubliés, dont nous pourrions tenir compte dans le redressement. Par exemple, on pourrait définir le rapport entre le revenu des ménages oubliés et le revenu moyen des ménages recensés ayant la même composition au moyen d'une régression linéaire; on pourrait ensuite imputer aux ménages additionnels (rajoutés) le revenu obtenu en appliquant la fonction de régression linéaire au revenu du ménage donneur recensé. Little et Rubin (1987) examinent des méthodes conçues pour résoudre les problèmes de données manquantes avec non-réponse informative. Dans la section suivante, nous décrivons une autre méthode intégrée à la méthode de redressement par pondération.

Le *redressement du nombre de personnes à l'intérieur de ménages* consiste à pondérer à la baisse un ménage ayant certaines caractéristiques recensées et à pondérer à la hausse un autre ménage qui compte un ou plusieurs membres additionnels. S'il n'y a pas d'autre redressement, ce sont les caractéristiques du ménage pondéré à la hausse, et non celles du ménage recensé qui a fourni le poids, qui s'appliqueront à la composante «rajoutée».

Cette situation pose des problèmes que l'on ne peut résoudre sans recueillir certaines données (dans un sous-échantillon de l'EP). Par exemple, si un *enfant* ne figure pas par erreur sur la liste des membres d'un ménage, il n'y a pas de raison de croire que cet oubli faussera le revenu du ménage. Si les ménages qui comptent relativement plus d'enfants avaient un revenu moyen plus élevé que celui des ménages ayant relativement moins d'enfants, la pondération tendrait à sur-estimer les revenus moyens.

Si un *adulte* était oublié, il serait permis de croire que le revenu de cet adulte (s'il en a un) ne serait pas inclus dans le revenu déclaré du ménage. Il est plausible de penser que le revenu moyen non déclaré dans un tel cas serait positif mais inférieur au revenu moyen d'un ménage de même composition, dont aucun des membres adultes n'aurait été oublié. Prenons l'exemple classique d'une famille qui vit d'aide sociale et ne déclare pas un adulte de sexe masculin, dont le domicile est plutôt instable et dont le revenu serait déduit du montant d'aide sociale si cette

Afin de satisfaire les utilisateurs qui aimeraient vérifier la sensibilité de leurs analyses au redressement pour sous-dénombrement, la bande devrait contenir des coefficients (notamment l'inverse des poids de redressement rattachés aux enregistrements de ménages qui figurent dans les listes originales du recensement) qui permettraient à l'utilisateur de reproduire les données non redressées du recensement.

8. REDRESSEMENT DES COVARIABLES QUI NE SERVENT PAS À LA CLASSIFICATION

Les méthodes décrites ci-dessus garantissent que les totaux pondérés d'îlots selon des variables classificatoires comme le sexe, l'origine ethnique et le groupe d'âge, égalent les totaux redressés d'îlots. Toutefois, ces listes serviront aussi à accumuler des totaux ou des chiffres de population pour des variables qui ne serviraient pas à la classification, par exemple le revenu et le niveau d'instruction. Dans cette section, nous examinons l'effet de ces méthodes de redressement sur des données de ce genre. Pour que l'illustration soit des plus concrètes, nous nous servirons du revenu comme exemple principal. Le revenu est une variable non classificatoire importante; des études permettent de croire que les programmes de répartition du revenu peuvent être très touchés par des erreurs de mesure du revenu. (National Academy of Sciences 1985).

En règle générale, il y a deux sources de biais possibles dans l'estimation d'une covariable non classificatoire: (1) biais dans le redressement de la composition du ménage, et (2) écarts systématiques entre les ménages recensés entièrement et les ménages de même composition qui ont été oubliés (en tout ou en partie). Cependant, si nous disposons d'une estimation du revenu moyen pour l'îlot, nous pouvons établir une relation d'égalité entre la moyenne pondérée pour les ménages de l'îlot et la moyenne estimée (redressée), de la même façon que nous le faisons pour le nombre pondéré de personnes dans l'îlot et leur nombre estimé (redressé).

8.1 Biais dû à la composition des ménages

Dans cette section, nous supposons que le niveau de revenu moyen qui se rattache à un certain mode de composition des ménages est le même pour les ménages qui ont été recensés entièrement et ceux qui ont été oubliés en tout ou en partie. Autrement dit, nous considérons ici le cas où le sous-dénombrement n'a aucun lien avec le revenu.

Supposons que le revenu du ménage est la somme des contributions indépendantes des membres du ménage répartis dans les diverses classes (autrement dit, supposons que la contribution des membres d'une classe est indépendante de la situation des autres membres dans le ménage). Le revenu total pondéré des ménages sera alors une estimation sans biais du revenu total réel (lorsque les taux de redressement sont exacts) puisque la somme des revenus est une fonction linéaire des effectifs de classe pour l'îlot. Toutefois, selon l'hypothèse plus réaliste de non-linéarité, une mauvaise répartition des personnes entre les ménages (et par conséquent une mauvaise représentation de la composition des ménages dans le redressement) pourrait introduire un biais dans les estimations du revenu. Ainsi par exemple, le revenu moyen des ménages qui comptent deux enfants pourrait ne pas correspondre à la moyenne des revenus moyens des ménages à un et à trois enfants (étant donné la même composition d'adultes). En conséquence, la pondération pourrait donner le bon nombre d'enfants mais si, en moyenne, on créait un trop grand nombre de ménages à deux enfants (comparativement à la réalité) par rapport au nombre de ménages à un ou à trois enfants, les estimations du revenu des ménages seraient biaisées.

Notre méthode tend à ajuster les poids qui donnent aux ménages «rajoutés» une composition semblable à celle des ménages recensés. Cependant, le redressement n'est décrit que par les totaux des classes de redressement, qui en disent peu sur la composition des ménages oubliés.

7. UTILISATIONS DE DONNÉES PONDÉRÉES

Le résultat des méthodes décrites dans les sections précédentes serait une liste de recensement dans laquelle les ménages auraient des poids, les membres de ces ménages auraient des poids identiques à celui du ménage et les pensionnaires d'établissement institutionnel auraient des poids qui leur auraient été attribués individuellement. Dans cette section, nous voyons comment ces listes servent aux fins du recensement.

7.1 Formation de tableaux de chiffres de population

Comme pour n'importe quelle série de données pondérées, la somme de poids remplace le nombre d'observations dans la création de tableaux. La seule difficulté liée à l'utilisation de poids est d'obtenir des nombres entiers dans les tableaux. Cette difficulté surgit même avant le calcul des poids des ménages: en effet, lorsqu'on estime le nombre de personnes ou de ménages oubliés, les effectifs des classes ne sont ordinairement pas exprimés en nombres entiers.

Si les totaux de classe redressés sont arrondis, un tableau qui fait la somme de plusieurs classes (par exemple, un nombre d'adultes de sexe masculin, qui est la somme des adultes de sexe masculin de diverses classes) contiendra aussi des entiers puisqu'il doit être conforme à ces totaux. En ce qui a trait aux tableaux qui ne reposent pas sur les totaux redressés, l'addition de poids dans un groupe particulier ne donnera pas nécessairement des nombres entiers. Par exemple, si une classe comprend des femmes de 20 à 40 ans, la somme des poids se rattachant aux femmes de 20 à 30 ans ne sera pas nécessairement un nombre entier. Quoi qu'il en soit, il serait peu vraisemblable d'arrondir tous les poids se rapportant à des classes puisque l'erreur d'arrondissement pourrait bien être du même ordre que les corrections. Toutefois, on devrait pouvoir se servir des méthodes d'arrondissement actuellement en usage au Census Bureau («arrondissement contrôlé») pour résoudre la difficulté, surtout lorsque les règles de protection du secret statistique exigent que les données destinées à la publication soient arrondies de toutes façons (Cox et coll. 1986; Cox 1987).

7.2 Formation de tableaux de sommes et de moyennes

Comme les sommes (de quantités continues) et les moyennes ne sont habituellement pas des nombres entiers, la question de l'arrondissement n'est pas pertinente ici. En outre, puisque les tableaux fondés sur les données de questionnaires complets proviennent déjà d'un échantillon, une source de poids additionnelle changerait peu de choses au processus. Ce à quoi il faut surtout s'intéresser ici sont les valeurs des covariables non classifiables qu'il faut attribuer aux ménages qui sont «rajoutés» dans le recensement; cette question est traitée dans la section 8.

7.3 Bandes-échantillon à grande diffusion

Les bandes à grande diffusion sont un échantillon d'enregistrements du recensement qui sont mis à la disposition des utilisateurs à des fins d'analyse. Pour produire ces bandes-échantillon à partir de listes pondérées du recensement, il suffit de modifier légèrement la méthode d'échantillonnage de manière que les probabilités d'échantillonnage soient proportionnelles aux poids. Même pour la bande à 5% (taux de sondage le plus élevé), les probabilités d'échantillonnage pondérées devraient être inférieures à 1. Une fois que ces bandes sont produites, l'utilisateur n'est pas obligé de connaître les méthodes de redressement et de pondération qui ont servi à la production de ces bandes. Les bandes à grande diffusion sont une source de données précieuse pour les analyses complexes réalisées par les sociologues, les économistes, les planificateurs, etc., et dans lesquelles la composition des ménages et les chiffres de population ont de l'importance. Il est essentiel que ces bandes puissent être produites facilement et utilisées comme des données brutes du recensement.

groupe de taille était celui qui avait été le plus sous-estimé par les données du recensement. Toutefois, étant donné la structure linéaire logarithmique du redressement, les corrections les plus importantes ont été apportées aux ménages ayant le plus de personnes et le moins de personnes. Par conséquent, les groupes de taille supérieurs ont été sur-redressés légèrement et les groupes intermédiaires ont été sous-redressés; le groupe des ménages de «taille 2» a été redressé légèrement dans la mauvaise direction. Néanmoins, la moyenne de la distribution redressée était beaucoup plus près de la valeur «réelle» que la moyenne redressée.

On observe la même chose dans le cas de la distribution du groupe d'âge de l'homme le plus âgé du ménage. Bien que ces données n'ont qu'un lien indirect avec les effectifs de classes, les distributions et les moyennes redressées reflètent mieux la «réalité» que les distributions et les moyennes non redressées dans presque tous les cas.

En définitive, ces simulations donnent à penser que le redressement par pondération peut améliorer les estimations des mesures de la structure des ménages de même que les données agrégées auxquelles il s'adresse. Toutefois, la repondération produit de moins bons résultats avec certaines formes de données, comme dans le cas des nombreux ménages qui comptent deux adultes; dans ce cas, on pourrait devoir utiliser une méthode d'imputation axée sur un modèle, comme celle décrite par Zaslavsky (1989).

Tableau 8

Résultats de la simulation d'inférence

Répartition des ménages selon la taille						
Répartition des ménages avec enfants selon la taille (nombre d'adultes)						
réel	recens	redress	taille 0	taille 1	taille 2	taille 3
7.240	10.349	7.372	16.200	19.631	20.240	22.600
3.971	3.632	3.971	33.720	27.558	34.219	39.720
Répartition des ménages avec enfants selon la taille (nombre d'adultes)						
réel	recens	redress	taille 0	taille 1	taille 2	taille 3
0.000	1.736	0.924	6.925	18.309	49.874	15.965
58.404	17.214	9.125	7.677	9.810	21.931	16.418
2.585	2.323	2.562	8.332	6.439	4.203	0.949
Répartition des ménages avec enfants selon la taille (nombre d'adultes)						
réel	recens	redress	taille 0	taille 1	taille 2	taille 3
7.080	9.981	7.853	4.000	7.388	26.296	30.972
21.960	21.160	21.931	4.480	4.203	4.480	0.949
Répartition des ménages avec enfants selon la taille (nombre d'adultes)						
réel	recens	redress	taille 0	taille 1	taille 2	taille 3
4.000	7.388	5.989	28.680	33.800	33.439	33.800
21.960	21.160	21.931	4.480	4.203	4.480	0.949
Répartition des ménages avec enfants selon la taille (nombre d'adultes)						
réel	recens	redress	taille 0	taille 1	taille 2	taille 3
3.602	5.809	4.272	6.214	11.723	27.242	42.038
15.843	15.158	16.418	0.949	0.894	0.962	2.638
Répartition des ménages avec enfants selon la taille (nombre d'adultes)						
réel	recens	redress	taille 0	taille 1	taille 2	taille 3
3.602	5.809	4.272	6.214	11.723	27.242	42.038
15.843	15.158	16.418	0.949	0.894	0.962	2.638
Répartition des ménages avec enfants selon la taille (nombre d'adultes)						
réel	recens	redress	taille 0	taille 1	taille 2	taille 3
3.602	5.809	4.272	6.214	11.723	27.242	42.038
15.843	15.158	16.418	0.949	0.894	0.962	2.638

6.2 Simulations d'inférence

Pour les simulations d'inférence, on a prélevé des pseudo-îlots de 50 ménages comptant uniquement des personnes d'origine hispanique. Ces pseudo-îlots ont été traités comme s'ils représentaient de vrais îlots. On a ensuite déterminé un niveau de sous-dénombrement simulé pour ces ménages en supposant que la probabilité qu'un membre du ménage soit oublié (indépendamment des autres) était égale au taux de sous-dénombrement calculé par Diffendal (1988), avec deux taux de sous-dénombrement négatifs ramenés à 0.

On a voulu représenter la distribution des compositions de l'îlot «recensé» en incluant dans la pseudo-liste de recensement la composition réelle des ménages et les compositions que l'on pouvait obtenir si un ou plusieurs membres du ménage étaient oubliés, chaque composition étant pondérée par la probabilité prévue dans le modèle.

On a ensuite répondu la pseudo-liste de recensement avec sous-dénombrement en fonction des totaux initiaux du pseudo-îlot pour le nombre de ménages et le nombre de personnes dans chaque classe de redressement. Enfin, on a comparé la pseudo-liste de recensement et la liste répondée au pseudo-îlot original.

La simulation a été effectuée ainsi de manière à éliminer la variabilité attribuable au caractère aléatoire du taux de sous-dénombrement d'un îlot (autour du taux de sous-dénombrement moyen) et de la répartition du nombre de personnes oubliées entre les ménages de l'îlot. En outre, l'applicabilité est garantie car les ménages initiaux sont toujours inclus (avec les poids) dans la pseudo-liste de recensement. L'interprétation que l'on peut donner de cela est que chaque îlot simulé représente une très grande population où les taux de sous-dénombrement observés et la distribution des compositions observées tendent vers leur espérance mathématique.

Plusieurs séries de données ont servi à analyser la méthode de ré pondération. Elles ont toutes été choisies parce qu'elles résument des caractéristiques du ménage qui ne sont pas fonction du nombre de personnes dans les classes de redressement. La première série de données concerne la distribution des tailles (nombre de membres) des ménages. Il convient de souligner que le nombre moyen de personnes par ménage, comme toute autre fonction des totaux de classe et du nombre de ménages, sera redressé automatiquement en fonction des valeurs exactes (avant sous-dénombrement); cependant, la distribution des tailles des ménages n'est pas soumise à ce redressement.

La seconde série de données portait sur la distribution du nombre d'*adultes* (plus de 14 ans) dans les ménages qui comptent un ou plusieurs *enfants* (14 ans et moins). Dans ce cas, la moyenne n'est pas redressée automatiquement en fonction de la valeur exacte puisqu'elle dépend des totaux marginaux et de la distribution conjointe de diverses classes à l'intérieur des ménages.

Les deux dernières séries de données portaient sur la distribution du groupe d'âge (codes de 1 à 5 comme pour la formation des classes de redressement) de l'*homme le plus âgé* du ménage (code 0 si le ménage ne compte pas d'homme) et sur la même distribution pour les ménages qui comptaient un ou plusieurs enfants. Là encore, ni la distribution ni sa moyenne ne correspondant directement à leur valeur réelle par contrainte.

Le tableau 8 donne un résumé des résultats de ces simulations. Comme presque tous les écarts indiqués ici sont très significatifs (par rapport aux variances inter-pseudo-îlot des écarts), les erreurs types ne figurent pas dans les tableaux. Dans chaque tableau, il y a une ligne «réel» (pour les pseudo-îlots originaux), une ligne «recens» (pour les îlots recensés simulés, c.-à-d. compte tenu des omissions dues au sous-dénombrement) et une ligne «redress» (îlots recensés, après redressement pour sous-dénombrement). Les chiffres de chaque colonne, sauf celle des moyennes, représentent le pourcentage de ménages dans l'îlot.

La distribution des tailles des ménages était biaisée à la baisse dans les îlots recensés. Le redressement a eu pour effet non seulement de corriger la moyenne mais aussi de rapprocher sensiblement le pourcentage estimé pour chaque taille du pourcentage réel.

La distribution du nombre d'*adultes* dans les ménages comptant un ou plusieurs enfants était aussi biaisée à la baisse. Comme la majorité de ces ménages comptaient deux adultes, ce

tableau 7 indique, pour chaque condition de simulation, la valeur moyenne de ces quantités (pour l'ensemble des flots simulés) dans les colonnes «Pmax», «Pmin» et «varP». Les observations ci-dessous décrivent quelques-uns des effets des facteurs du plan de simulation sur les poids ajustés.

(1) L'utilisation d'un facteur de redressement du nombre de ménages de 1,05 (par rapport à un facteur de (1) donne dans chaque cas une variance moyenne des poids inférieure et une moyenne des poids minimum et des poids maximum plus près de 1. Ces résultats sont acceptables intuitivement car presque tous les facteurs de redressement pour les classes sont supérieurs à 1 et qu'il faut procéder à un redressement plus prononcé lorsqu'on ajoute des personnes dans des ménages existants que lorsqu'on ajoute simplement des personnes et des ménages pour les recevoir. Par exemple, si les facteurs de redressement pour les ménages et chaque classe de redressement étaient tous égaux, chaque ménage serait pondéré à la hausse également.

(2) Les autres facteurs étant fixes, la variance des poids diminue à mesure qu'augmente le nombre de ménages par flot. Ce résultat est aussi acceptable intuitivement car le groupe des ménages est plus diversifié dans un flot plus grand; la probabilité de trouver exactement les ménages voulus pour représenter les personnes qui n'ont pas été dénombrées est plus élevée. Les tendances des poids extrêmes se dégagent moins clairement que celles des variances; ici, la réduction de la variance est contrebalancée par l'échantillon plus grand par rapport auquel est calculé le poids extrême dans les flots plus grands.

(3) En ce qui a trait aux simulations concernant les flots de 200 ménages, la variance moyenne n'a pas dépassé 0,063. Il est donc permis de conclure qu'en règle générale, la reproduction ne produit pas de poids extrême.

Coût des calculs

Le nombre moyen d'itérations de Newton nécessaires pour ajuster les poids (sur la base des valeurs initiales définies dans la section 3.3) est habituellement de 2; le nombre d'itérations figure dans la colonne «itér» du tableau 7. Deux itérations ont suffi pour satisfaire toutes les contraintes avec une marge d'erreur n'excédant pas 0,001. À l'aide de cette information, il est possible d'estimer grossièrement le nombre d'opérations en virgule flottante nécessaires pour appliquer l'algorithme. Le coût d'application de l'algorithme d'itération modifié est analysé dans la section 10.2.

Supposons que les flots sont de taille suffisamment grande pour ne pas avoir à vérifier la cohérence et l'applicabilité des contraintes dans chaque cas (sauf peut-être lorsque l'ajustement de poids ne peut être accompli en quelques itérations). L'opération essentielle est donc d'ajuster les poids. Pour les phases d'exécution, les programmes et les structures de données devraient être conçus de manière à tirer profit de la faible densité de la matrice A (qui tient au fait que seules quelques classes sont représentées dans chaque ménage). Si S_1 est le nombre total d'éléments non nuls dans A et S_2 est la somme (pour l'flot) des *carres* du nombre d'éléments non nuls pour chaque ménage, chaque itération de Newton exige environ $S_1/2 + S_2/2$ multiplications (plus un terme indépendant du nombre de ménages par flot). Dans les échantillons étudiés ici, $S_2 \approx 5S_1$; S_1 est borné par la population totale de l'flot. Par conséquent, le nombre maximum de multiplications est environ de $15 \times$ total de la population (si l'on compte l'étape de départ comme une itération); le nombre maximum d'additions est comparable.

À une époque où même les micro-ordinateurs ont une puissance arithmétique qui se mesure en mégaflops, il ne semble pas déraisonnable d'exécuter 8×10^9 opérations en virgule flottante pour répondre les résultats d'un recensement tout entier. Le calcul des poids pourrait exiger moins de ressources que le traitement de données «administratives», qui accompagne toute méthode d'analyse du sous-dénombrement. Evidemment, si la méthode était appliquée à un échantillon de base de données, comme une bande-échantillon à grande diffusion, le coût baisserait en conséquence.

Elles étaient habituellement applicables pour ce qui est des flots de 100 ménages et toujours applicables dans le cas des flots de 200 ménages.

Les colonnes numérotées à droite représentent l'ordre de la plus simple contrainte marginale qui n'a pu être satisfaite au sens de la reparamétrisation hiérarchique définie dans la section 5.2.1. Ainsi, la colonne (1) indique le nombre d'flots simulés pour lesquels une contrainte d'«effet principal» (total marginal des personnes classées selon une variable de stratification) n'a pu être satisfaite, la colonne (2) indique le nombre d'essais pour lesquels une contrainte d'interaction à deux critères n'a pu être satisfaite, etc. Même lorsque les contraintes étaient incohérentes dans le cas d'flots de 50 ou 100 ménages, les contraintes d'effet principal et, souvent, les contraintes d'interaction à deux ou même à trois critères étaient applicables. Cela donne à penser que le groupement d'flots pour des interactions d'ordre supérieur, comme nous l'avons vu dans la section 5.2.3, pourrait être une solution efficace aux problèmes d'inapplicabilité. Les résultats ont été moins encourageants dans le cas des simulations qui portaient sur les échantillons complets. Même pour des flots de 200 ménages, les contraintes étaient rarement cohérentes et applicables. À mesure qu'augmentait la taille de l'flot, les contraintes d'ordre inférieur étaient plus susceptibles d'être applicables. Cela s'explique par le faible nombre de ménages qui comptent des personnes d'origine asiatique (environ 5 % dans chaque échantillon). Sur 200 ménages, on s'attendrait à en avoir environ 10 qui comptent des personnes d'origine asiatique, ce qui est insuffisant pour satisfaire les 20 contraintes possibles pour les classes de redressement relatives aux personnes d'origine asiatique. Dans la réalité, il n'est sûrement pas rare de constater que des groupes de classes de redressement sont mal représentés dans une région ou des flots en particulier. Il faudra alors procéder à un groupement d'flots à grande échelle pour les contraintes correspondantes tout en essayant de satisfaire à une plus petite échelle les contraintes se rapportant aux classes mieux représentées.

Poids:

Les poids maximum et minimum des ménages de même que leur variance ont été calculés pour chaque flot simulé pour lequel les contraintes étaient cohérentes et applicables. Le

Tableau 7

Résultats de la simulation concernant l'applicabilité

Ménages ne comptant aucune personne d'origine asiatique											
Taille	HH	Taux	incohér	inap	OK	Pmax	Pmin	varP	itér	(1)	(2)
10	10	1.00	50	0	0	NA	NA	NA	NA	22	28
10	10	1.05	50	0	0	NA	NA	NA	NA	8	42
20	20	1.00	50	0	0	NA	NA	NA	NA	0	50
20	20	1.05	50	0	0	NA	NA	NA	NA	0	50
50	50	1.00	47	1	2	1.921	0.200	0.142	3.00	0	3
50	50	1.05	47	0	3	1.550	0.620	0.036	1.33	0	3
100	100	1.00	10	0	40	2.068	0.429	0.088	2.03	0	0
100	100	1.05	10	0	40	1.573	0.753	0.020	1.90	0	0
200	200	1.00	0	0	50	2.434	0.543	0.063	2.18	0	0
200	200	1.05	0	0	50	1.749	0.821	0.015	2.00	0	0
Échantillon complet											
Taille	HH	Taux	incohér	inap	OK	Pmax	Pmin	varP	itér	(1)	(2)
100	100	1.00	0	1	1	--	--	--	--	0	34
100	100	1.00	49	0	1	--	--	--	--	0	2
200	200	1.00	49	0	1	--	--	--	--	0	43
200	200	1.00	49	0	1	--	--	--	--	0	4

Pour répondre à ces questions, nous avons formé des flots simulés constitués de vrais ménages pour représenter les modes de composition réels (mais inobservés) des ménages dans les flots. Pour chaque «vrai» flot, nous avons fixé un niveau de sous-dénombrement à l'aide de taux de sous-dénombrement réels estimés et d'un modèle plausible pour la répartition du sous-dénombrement entre les ménages. Nous avons appliqué l'algorithme de pondération aux flots «recensés» ainsi obtenus. Des données sommatrices sur la composition des ménages ont été calculées pour les flots simulés «réels» et les flots simulés observés où il y avait sous-dénombrement; il s'agissait aussi bien de données non pondérées que de données pondérées pour le redressement en fonction du sous-dénombrement. Ces «simulations d'inférence» avaient pour but de déterminer si la reproduction rapprochait les données précitées des valeurs correspondantes dans les flots «réels»; autrement dit, la reproduction corrigeait-elle les biais attribuables au sous-dénombrement?

La source des ménages pour toutes les simulations a été l'échantillon «B» de microdonnées à grande diffusion de 1 % (Public Use Microdata Sample -- PUMS) du recensement de 1980 (Bureau of the Census 1985). Les ménages ont été prélevés dans des secteurs du Los Angeles County en Californie, parmi lesquels se trouve la région ayant servi au Test des opérations de redressement (TOR) du recensement d'essai de 1986. Les taux de sous-dénombrement étaient ceux qui avaient été calculés dans le TOR de 1986 (Diffendal 1988; tableau 7) pour des classes de redressement définies selon le sexe, l'âge (5 niveaux), l'origine ethnique (hispanique, asiatique ou autres) et le mode d'occupation (proprétaire ou locataire). Les facteurs de redressement calculés à partir de ces taux de sous-dénombrement variaient de 0,982 à 1,211. Chaque ménage était codé comme un vecteur de chiffres représentant le nombre de membres du ménage qui appartenaient à chacune des 60 classes de redressement. Zaslavsky (1989) donne des renseignements plus détaillés sur les méthodes de simulation et une autre série de simulations plus vaste.

6.1 Simulations concernant l'applicabilité

Pour l'une et l'autre de quatre tailles de flot (20, 50, 100 et 200 ménages), on a prélevé 50 flots simulés dans l'échantillon complet et on en a formé 50 autres avec les ménages qui ne comptaient aucune personne d'origine asiatique. Pour chaque flot, les simulations ont été effectuées avec deux valeurs du taux de redressement des ménages (facteur par lequel est redressé le nombre de ménages dans l'flot). On a appliqué les algorithmes de la section 3. En résumé, on a tout d'abord vérifié la cohérence des contraintes linéaires, puis leur applicabilité (existence d'une solution positive); enfin, on a calculé les poids à l'aide de la méthode de Newton. Comme il n'y avait pas de données qui permettaient de distinguer le sous-dénombrement des ménages du sous-dénombrement à l'intérieur des ménages, on n'a pas tenté de le faire dans ces simulations ou dans d'autres. Les résultats des simulations sont résumés dans le tableau 7.

Cohérence et applicabilité

Les colonnes identifiées «incohér», «inap» et «OK» indiquent le nombre d'flots simulés (sur les 50 essais) dans chaque simulation, qui faisaient partie de l'une ou l'autre des catégories suivantes: (1) les contraintes étaient incohérentes (ne pouvaient être satisfaites par aucun poids); (2) les contraintes étaient cohérentes mais inapplicables (ne pouvaient être satisfaites par aucun poids positif), ou (3) les contraintes étaient à la fois cohérentes et applicables. Dans les simulations portant sur les flots où il n'y avait aucune personne d'origine asiatique, il fallait satisfaire 41 contraintes (dont certaines pouvaient être banales, c'est-à-dire lorsque les classes de redressement correspondent à des flots). Ainsi, pour ce qui est des flots de 20 ménages, les contraintes n'étaient jamais cohérentes; dans le cas des flots de 50 ménages, les contraintes étaient parfois cohérentes et, en règle générale, applicables.

observés s'il y avait eu dénombrement complet. On fait le lissage de la distribution observée des modes de composition des ménages en la combinant à la distribution observée pour des îlots adjacents, qui renferment des ménages tout aussi caractéristiques pour ce secteur. Cette méthode se rattache donc théoriquement aux méthodes de lissage bayésiennes qui améliorent l'estimation d'une quantité pour une unité en tirant parti de la distribution observée dans des unités semblables. Zaslavsky (1989) intègre ce raisonnement bayésien dans un modèle d'effets aléatoires au niveau de l'îlot.

On pourrait choisir les îlots donneurs par une méthode hot-deck récurative; ces îlots pourraient le plus souvent se trouver près de l'îlot de redressement et aucune série d'îlots en particulier n'aurait une influence anormale sur l'ensemble du recensement. Par une stratification détaillée des îlots, on pourrait faire en sorte de choisir des îlots donneurs qui ressembleraient à l'îlot visé par le redressement au point de vue du revenu moyen, des catégories d'unités de logement et de la répartition raciale.

5.2.3 Combinaison de méthodes

Les deux catégories de méthodes exposées ci-dessus peuvent être combinées par une redéfinition appropriée des contraintes. Le principe, en l'occurrence, est de satisfaire *toutes* les contraintes dans les grandes unités géographiques et de ne satisfaire que les plus importantes dans les petites unités. Ce genre de compromis peut permettre d'obtenir un ajustement assez valide de la distribution voulue sans devoir allonger la liste des ménages.

Supposons que les matrices A de plusieurs îlots ont toutes été réécrites comme des suites de lignes représentant les contraintes principales et les contraintes d'interaction. On peut alors former une seule grande matrice A qui représente toutes les contraintes. Les lignes qui représentent les contraintes principales peuvent demeurer telles quelles tandis que les lignes qui représentent les contraintes secondaires peuvent être combinées dans les îlots. Supposons, par exemple, qu'il y a dix classes de redressement, définies selon le sexe (2 niveaux) et l'âge (5 niveaux), et deux îlots. Il y a en tout 11 contraintes (une pour le nombre de ménages et une pour chaque classe de redressement) dans chaque îlot. Si on combine ces contraintes dans une seule matrice en conservant les effets principaux et les interactions à deux critères, on a les contraintes suivantes: nombre de ménages dans l'îlot (2 contraintes), population de l'îlot (2 contraintes), sexe (1 contrainte), âge (4 contraintes), interaction îlot et sexe (1 contrainte), interaction îlot et âge (4 contraintes) et interaction îlot \times sexe \times âge; dans un exemple plus réaliste, où il y aurait un plus grand nombre d'îlots, de variables de classification et de niveaux, on pourrait en éliminer beaucoup plus.

6. RÉSULTATS DE SIMULATIONS

Des simulations ont été faites pour répondre à deux séries de questions:

- (1) La première série de questions vise à évaluer l'efficacité de l'algorithme en fonction de ses contraintes et de ses objectifs. L'algorithme de pondération donne-t-il une réponse? Dans les problèmes réels, y a-t-il une solution aux contraintes de pondération? De combien les poids varient-ils? La quantité de calculs nécessaires est-elle raisonnable?
- (2) La seconde série de questions vise à évaluer dans quelle mesure l'algorithme améliore la qualité des inférences fondées sur une série de microdonnées: la série de microdonnées pondérées décrit-elle mieux la réalité que les données brutes, non pondérées?

Pour répondre à ces questions, nous avons procédé à des «simulations d'applicabilité», où nous avons appliqué l'algorithme de pondération à des îlots simulés composés de vrais ménages en utilisant des taux de redressement réels. Cette méthode reproduit l'application de l'algorithme.

ou incohérentes. En extrayant simplement ces lignes de la matrice A , on obtient un ensemble cohérent de contraintes et la s'arrêtent les calculs requis.

Si les contraintes sont disposées par ordre décroissant d'importance, les moins importantes sont éliminées si elles sont incompatibles avec les plus importantes. Cette disposition est tout à fait logique si les contraintes initiales portant sur les classes de redressement (définies par une classification de la population selon plusieurs critères) sont redéfinies selon les règles de l'analyse de variance des contraintes concernant la population totale («moyenne globale»), les classes définies par une variable de classification («effets principaux») et les classes définies par des interactions. Par exemple, si nous avons 10 classes de redressement définies par le sexe et cinq groupes d'âge, les nouvelles contraintes seraient par ordre d'importance: population totale (1 contrainte), population selon le sexe (1 contrainte additionnelle), population selon l'âge (4 contraintes additionnelles), interactions âge-sexe (les 4 dernières contraintes). Les quatre contraintes fondées sur l'âge pourraient être subdivisées entre personnes jeunes et personnes âgées (1 contrainte) et 3 contraintes additionnelles pourraient être définies pour chacun de ces groupes.

On peut appliquer une méthode similaire au moment de vérifier l'applicabilité des contraintes. Si il n'est pas possible de satisfaire l'équation $Z_i = 0$, pour tous les i , on peut modifier la fonction objectif du problème de programmation linéaire de manière à obtenir $\sum c_i Z_i$, où les coefficients $c_i > 0$ correspondent aux contraintes les plus importantes prennent les valeurs les plus grandes. On peut alors définir un ensemble maximal de contraintes applicables et supprimer les autres contraintes.

De cette façon, on obtient des poids qui donnent des totaux d'îlots justes pour les classes de personnes plus générales mais non pour tous les tableaux croisés.

5.2.2 Méthodes fondées sur l'addition de colonnes (ménages) à A

Lorsque l'inalplicabilité des contraintes n'est que conditionnelle (c'est-à-dire qu'elle dépend des modes de composition des ménages de l'îlot), il suffit d'ajouter des ménages qui ont la composition voulue pour rendre les contraintes applicables. Le moyen le plus simple d'appliquer ce principe est de travailler à un niveau d'aggrégation géographique supérieur à celui de l'îlot. Lorsque des difficultés surgissent dans l'ajustement, on peut combiner quelques îlots adjacents ou former une liste au niveau du secteur de dénombrement par exemple, avant la pondération. Plus l'unité sera grande, plus les modes de composition des ménages représentés seront nombreux et moins il y aura de chances d'observer des problèmes d'inalplicabilité.

Une méthode plus poussée consisterait à utiliser un échantillon de ménages prélevé par hot-deck dans des îlots «donneurs» adjacents afin d'élargir le groupe de ménages auxquels un poids peut être attribué. Il importe ici de s'en tenir à un calcul simple puisqu'on pourrait devoir consulter une longue liste de ménages afin de trouver le ou les ménages qui rendront les contraintes applicables. Pour ce qui a trait à la vérification de la cohérence des contraintes, si la ligne j de A est dépendante des lignes précédentes et que la colonne correspondante aux ménages ajoutés est indépendante des colonnes de A (en ce qui a trait uniquement aux j premières lignes), la ligne j de la matrice A enrichie sera indépendante. Pour ce qui a trait à la vérification de l'applicabilité, si l'algorithme est interrompu parce qu'aucune réduction n'est possible dans la fonction objectif $\sum Z_i$, on peut chercher des colonnes de base parmi les colonnes correspondant aux ménages de l'échantillon prélevé par hot-deck. Enfin, si le poids ajusté d'un ménage est extrêmement élevé, on peut chercher parmi les ménages de l'échantillon hot-deck des ménages auxquels on attribuerait aussi des poids élevés avec les valeurs courantes de (c'est-à-dire, colonnes a de telle sorte que $a' \lambda$ est élevé). Si ces ménages sont ajoutés aux autres ménages de l'îlot, ils prendront une partie du poids des ménages surpondérés lorsque les poids seront réajustés puisqu'ils sont susceptibles de compter aussi des membres des mêmes classes de redressement.

Le raisonnement intuitif sur lequel repose cette méthode est que les modes de composition qui ont été observés dans un îlot ne sont qu'un échantillon des modes qui auraient pu y être

faillie pour cela redresser à outrance le poids de certains ménages. Les problèmes liés à ces trois situations sont relativement comparables.

Il est possible de définir des contraintes qui sont incohérentes en soi, par exemple que toutes les classes d'hommes sont redressées de 2% à la hausse tandis que le nombre total d'hommes est redressé de 4% à la hausse. Selon la méthode que nous utilisons, chaque contrainte porte uniquement sur le nombre de personnes contenues dans une classe de redressement de sorte qu'il n'y a pas d'incohérence du genre de celle mentionnée ci-dessus. Toutefois, il y a encore la possibilité d'une incohérence conditionnelle, c'est-à-dire d'une incohérence qui dépend du nombre de ménages correspondant à chaque mode de composition dans un îlot. Nous donnons ci-dessous des exemples d'incohérence conditionnelle, d'inapplicabilité ou de poids insatisfaisants:

- (1) Selon les méthodes proposées pour estimer le taux de sous-dénombrement, il est question de définir plus de 100 classes de redressement. Dans un îlot restreint mais hétérogène, le nombre de classes représentées pourrait être supérieur au nombre de ménages; par conséquent, il y aurait plus de contraintes que de poids à ajuster. Une incohérence est alors presque inévitable.
- (2) Si tous les ménages d'un îlot comptent exactement le même nombre de personnes d'une classe de redressement en particulier (par ex., chaque ménage compte une jeune fille d'origine hispanique), le nombre de membres de cette classe n'est pas modifié par la redistribution des poids.

- (3) Le redressement du nombre de ménages peut aboutir à des résultats incompatibles avec ceux qui découleraient du redressement du nombre de personnes dans une classe quelconque (cela pouvant être interprété comme une défaillance du modèle pour le redressement du nombre de ménages). Prenons par exemple le cas où il faudrait ajouter plus d'hommes que de ménages mais où aucun ménage de l'îlot ne compte plus d'un homme. Les contraintes seraient alors cohérentes mais inapplicables puisqu'on ne pourrait les satisfaire qu'en attribuant des poids négatifs aux ménages qui ne comptent pas d'hommes.

- (4) Un îlot peut avoir des taux de sous-dénombrement qui diffèrent totalement de ceux qui ont été estimés pour des îlots de l'EP. Supposons, par exemple, qu'il y a un fort taux de sous-dénombrement dans la plupart des îlots (y compris la plupart des îlots de l'échantillon de l'EP) pour les hommes qui présentent certaines caractéristiques mais que dans l'îlot faisant l'objet d'un redressement, les hommes de cette catégorie sont bien dénombrés et se retrouvent dans la plupart des ménages. L'estimation du niveau de sous-dénombrement pour cette classe pourrait entraîner un tel redressement à la hausse qu'il ne serait pas possible de concilier les résultats de ce redressement avec les chiffres des ménages existants.

- (5) On pourrait devoir hausser sensiblement le poids des ménages qui comptent des personnes issues d'une combinaison de classes de redressement propre à un ménage, de sorte que les ménages en question recevraient des poids extrêmes. Dans ce cas, le problème peut être résolu mais la solution n'est pas très satisfaisante.

Des problèmes d'inapplicabilité peuvent également surgir lorsqu'on ne peut associer aussi facilement la difficulté à une incohérence particulière dans le processus de redressement.

5.2 Méthodes pour rendre les contraintes applicables

Peu importe l'étape de l'ajustement à laquelle on constate l'inapplicabilité, il existe plusieurs méthodes permettant d'assouplir les contraintes et de les rendre applicables. Nous allons maintenant analyser plusieurs de ces méthodes en exposant la logique intuitive qui est à l'origine du choix de chacune d'elles de même que les méthodes de calcul requises.

5.2.1 Méthodes fondées sur l'élimination de lignes (contraintes) de A

Lorsqu'on vérifie la cohérence des contraintes, on peut découvrir que certaines lignes sont linéairement dépendantes des lignes précédentes et que, par conséquent, elles sont superflues

Tableau 6

Nombre hypothétique de ménages pour l'exemple 3 (chiffres bruts et chiffres redressés)

Composition du ménage		(1) Les personnes oubliées font partie de ménages recensés ou non		(2) Les personnes oubliées font partie de ménages non recensés	
Nombre de personnes de la classe 1	Nombre de personnes de la classe 2	Chiffres bruts (nombre de ménages)	Chiffres redressés	Nombre de ménages non recensés	Totaux redressés, ménages recensés et non recensés
1	1	10,000	9904,54	.01	10,000.01
1	2	10,000	10106,46	10,99	10,010.99
2	1	10,000	10106,46	10,99	10,010.99
2	2	10	13,54	99,01	109,01

Supposons que nous devons ajouter 231 personnes de chaque classe et 121 ménages aux 30,010 ménages recensés. Les trois dernières colonnes du tableau 6 contiennent les chiffres redressés selon deux hypothèses: (1) les personnes oubliées peuvent faire partie de ménages qui ont été recensés ou non et (2) toutes les personnes oubliées font partie des ménages qui n'ont pas été recensés.

Dans le cas de la première hypothèse, l'algorithme pondère à la baisse les ménages qui ne comptent qu'une personne de chaque classe (1,1) et pondère à la hausse les ménages qui comptent deux personnes d'une classe et une de l'autre (1,2 et 2,1). Même si les ménages qui comptent deux personnes de chaque classe sont fortement pondérés à la hausse (par un facteur de 1,354), ils ne reçoivent qu'une faible proportion des personnes ajoutées à cause de leur très petit nombre à l'origine.

Dans le cas de la seconde hypothèse, on calcule tout d'abord des poids de manière à intégrer $231 \times 2 = 462$ personnes dans 121 ménages additionnels, puis on additionne ces poids au poids unitaire contenu dans les chiffres bruts. Tandis qu'aucune catégorie de ménages n'est pondérée à la baisse, les ménages qui comptent deux personnes de chaque classe (2,2) sont très fortement pondérés à la hausse (par un facteur de 10,901). De fait, il est mathématiquement impossible d'intégrer 462 personnes dans 121 ménages formés de deux, trois ou quatre personnes chacun sans avoir au moins 99 ménages de quatre personnes. Par conséquent, le fait de savoir que les personnes ajoutées (ou une proportion connue de celle-ci) font partie des ménages qui ont été oubliés modifie sensiblement notre représentation du genre de redressement à effectuer.

5. APPLICABILITÉ DES CONTRAINTES

Dans les sections précédentes, nous avons supposé qu'il y avait des solutions réalisables pour le problème d'optimisation conditionnelle. Nous allons maintenant considérer des cas où les solutions sont inexistantes ou insatisfaisantes et allons examiner des méthodes pour résoudre ces cas.

5.1 Dans quels cas les contraintes sont-elles inapplicables?

Il y a trois cas où les contraintes peuvent éliminer toute possibilité de solution satisfaisante: (1) lorsqu'elles sont effectivement incohérentes, (2) lorsqu'elles sont cohérentes mais qu'il n'y a aucun poids positif pour les satisfaire, et (3) lorsqu'il y a une solution réalisable mais qu'il

Exemple 2: (suite).

Le tableau 5 donne le poids par ménage et le total pondéré des ménages (poids multiplié par le chiffre brut) pour chaque ligne du tableau 3. Aucun ménage n'est pondéré de plus de 8% à la hausse et de plus de 5% à la baisse.

4. REDRESSEMENT DU NOMBRE DE MÉNAGES
ET DU NOMBRE DE PERSONNES À
L'INTÉRIEUR DES MÉNAGES

Nous voyons maintenant la distinction à faire entre le redressement du nombre de personnes à l'intérieur des ménages (c'est-à-dire le redressement en fonction du sous-dénombrement à l'intérieur des ménages) et le redressement du nombre de ménages (c'est-à-dire le redressement en fonction du sous-dénombrement des ménages). Cette distinction avait déjà été faite en vue d'analyser les causes du sous-dénombrement (Fay 1986). Nous voulons ici nous en servir pour représenter plus fidèlement le sous-dénombrement dans une opération de redressement. Le redressement du nombre de personnes à l'intérieur des ménages ne consiste pas à ajouter des ménages à la liste mais uniquement à redistribuer les poids des ménages de manière à accroître les totaux pondérés de personnes dans les diverses classes. En d'autres termes, les ménages qui n'ont personne ou qui comptent très peu de personnes dans une classe donnée sont pondérés à la baisse tandis que ceux qui en comptent beaucoup sont pondérés à la hausse de manière que le poids total des ménages demeure constant. Ainsi, ce mode de redressement aura pour conséquence inéluctable de réduire le poids de certains ménages. En revanche, le redressement du nombre de ménages touche les ménages qui n'ont pas été du tout recensés. Comme ce mode de redressement ne doit pas modifier les données relatives aux ménages recensés, il convient d'ajouter des ménages à la liste sans réduire le poids des ménages qui ont été recensés.

Nous proposons de distinguer ces deux modes de redressement. Une série de contraintes représente le redressement du nombre de personnes à l'intérieur des ménages. Dans ce cas, le poids total des ménages doit être égal au nombre de ménages recensés tandis que le poids total des personnes dans chaque classe doit être égal à la somme du nombre de personnes recensées et du nombre redressé de personnes pour cette classe. $AW_1 = B_1$, où B_1 est constituée du nombre de ménages recensés et des effectifs de classes redressés en fonction du sous-dénombrement à l'intérieur des ménages.

Une seconde série de contraintes représente le redressement du nombre de ménages. Dans ce cas, le poids total des ménages doit être égal au nombre estimé de ménages oubliés tandis que le poids total des personnes dans chaque classe doit être égal au nombre estimé de personnes dans les ménages oubliés. $AW_2 = B_2$, où B_2 est constituée du nombre de ménages additionnels et du nombre de personnes additionnelles par classe, redressés en fonction du sous-dénombrement des ménages.

Après avoir ajusté deux séries de poids correspondant aux deux séries de contraintes, nous additionnons les deux poids pour chaque ménage afin d'obtenir un poids qui intègre les deux modes de redressement ($W = W_1 + W_2$). La différenciation des deux modes de redressement peut produire une série de poids redressés différente de celle qui serait obtenue si on ne faisait pas une telle distinction, comme le montre l'exemple 3. Toutefois, si on ne fait pas cette distinction dans l'estimation du taux de sous-dénombrement, il est quand même possible d'opérer un redressement en une seule étape.

Exemple 3: redressement en fonction du sous-dénombrement des ménages.

Supposons qu'il n'y a que deux classes de redressement et qu'un flot hypothétique est composé de la façon décrite par les trois premières colonnes du tableau 6.

Tableau 5
Poids optimaux pour l'exemple 2

Ligne n°	Poids	Chiffres pondérés
1	0.9554	47.77
2	0.9557	38.23
3	0.9816	39.27
4	0.9823	14.73
5	1.0730	53.65
6	1.0734	64.40
7	1.0737	42.95

3.2 Existence de solutions réalisables

Déterminer l'existence de solutions réalisables équivaut à déterminer une solution réalisable initiale dans un problème de programmation linéaire où les algorithmes réguliers peuvent être utilisés. Supposons que nous devons trouver une solution positive W à l'équation $AW = B$, où $B \geq 0$. (Si cette dernière condition n'est pas respectée, nous pouvons faire en sorte qu'elle le soit en modifiant le signe des éléments négatifs de B et des lignes correspondantes de A) Nous pouvons élargir le problème en posant $[A \mid I] \begin{bmatrix} W' \\ Z' \end{bmatrix} = B, W, Z \geq 0$, où I est une matrice unité $p \times p$ et Z est une variable vectorielle à p éléments. Ce problème a automatiquement une solution initiale $W = 0, Z = B$. Nous appliquons ensuite la méthode du simplexe (comme dans Gass (1964) ou tout autre ouvrage sur la programmation linéaire) pour minimiser $\sum Z_i$. Si cette somme peut être réduite à 0, les valeurs W correspondantes sont une solution au problème original, tandis que si elle ne peut l'être, le problème original n'a pas de solution.

Exemple 2: (suite).

Une solution réalisable (mais non optimale) donne des totaux pondérés de 86, 54, 29 et 132 pour les ménages des lignes 2, 3, 5 et 6 respectivement dans le tableau 3. Il est possible de vérifier que ces chiffres donnent les totaux redressés prévus pour les ménages et les personnes de chaque classe.

Le problème de l'inapplicabilité se rapproche de celui de l'incohérence et est aussi analysé dans la section 5.

3.3 Optimatisation de la fonction objectif

Selon la méthode des multiplicateurs de Lagrange, la solution de minimisation doit satisfaire les équations $\partial T / \partial W_i = \log W_i + 1 = a_i' \lambda$, où a_i' représente la i -ème colonne de A et $\lambda' = (\lambda_1, \lambda_2, \dots, \lambda_p)$. Alors, $W_i' = \exp(a_i' \lambda - 1)$; le modèle pour les poids est donc de forme logarithmique linéaire, comme celui utilisé pour les redressements par la méthode itérative du quotient. λ_s représente l'accroissement de poids logarithmique lié à un accroissement unitaire du coefficient de liaison correspondant a_{is} , par exemple, l'inclusion dans le ménage d'une personne faisant partie de la classe de redressements.

Pour satisfaire l'équation $AW = B$, nous pouvons résoudre en fonction de λ à l'aide de la méthode de Newton. Le modèle d'itération utilisé est

$$\lambda^{(t+1)} = \lambda^{(t)} - (AW^* A')^{-1} (AW - B), \tag{4}$$

où W est la matrice dont la diagonale est formée des éléments de $W = W(\lambda^{(t)})$. Une bonne valeur initiale pour λ est $\lambda^{(0)} = (A A')^{-1} B$, laquelle peut être déduite d'une approximation linéaire fondée sur des poids initiaux égaux. (Dans la section 10.2, nous décrivons une méthode de la descente cyclique pour résoudre ces équations; cette méthode est une généralisation de l'ajustement proportionnel itératif).

Tableau 3

Liste de ménage				
Nombre de membres du ménage par classe				
Ligne n°	Classe 1 (hommes)	Class 2 (femmes)	Class 3 (enfants)	Nombre de ménages
1	0	1	0	50
2	0	1	1	40
3	1	0	0	40
4	1	0	2	15
5	1	1	0	50
6	1	1	1	60
7	1	1	2	40

Tableau 4

Totaux redressées			
Chiffre brut	Taux de redressement	Chiffre redressé	
Classe 1 205	.05	215	
Classe 2 240	.03	247	
Classe 3 210	.04	218	
Ménages 295	.02	301	

Exemple 2: Ajustement de poids.

Le tableau 3 est un exemple d'une liste des ménages d'un îlot, où trois classes sont représentées comme dans l'exemple 1; nous pouvons penser à des classes comme «hommes», «femmes» et «enfants». Ce tableau peut être considéré comme la version condensée d'un tableau de 295 lignes, où chaque ligne correspond à un ménage.

Le tableau 4 contient le nombre non redressé et le nombre redressé de ménages et de personnes dans chaque classe. On obtient le nombre redressé en appliquant le taux de redressement indiqué et en arrondissant. La méthode utilisée pour obtenir les chiffres redressés n'a toutefois rien à voir avec le reste du processus.

3.1 Cohérence des contraintes linéaires

Tant que les lignes de A sont indépendantes, les contraintes $AW = B$ seront cohérentes. Si une ligne quelconque de A est dépendante des autres, la contrainte correspondante est soit incohérente ou superflue, selon les valeurs de B . On peut repérer les lignes dépendantes en appliquant la décomposition $\bar{Q}-R$ à A' . Si les contraintes correspondantes sont superflues, on peut les supprimer sans perte d'information; si elles sont incohérentes, elles doivent être reformulées de quelconque façon.

Exemple 2: (suite).

Dans cet exemple, les lignes de la matrice A sont indépendantes et les contraintes sont donc cohérentes.

Dans la section 5, nous examinons des situations où il est susceptible d'exister des contraintes incohérentes et analysons des méthodes qui permettent de les résoudre.

un prolongement de la méthode itérative du quotient. Scheuren (1973) utilise la méthode itérative du quotient pour la ré pondération des ménages; Cilke et Wyscarver (1988) font une ré pondération en fonction de contraintes linéaires mais utilisent pour cela une fonction objectif différente de celles considérées ici. Alexander (1987) a élaboré de son côté des méthodes semblables à celles exposées ici.

Pour ce qui a trait à la méthode itérative du quotient, des fréquences initiales X figurent dans les cases d'un tableau de contingence et on calcule de nouvelles fréquences de case Y pour minimiser la fonction objectif $\sum Y_i \log (Y_i/X_i)$. Par conséquent, les poids des observations originales sont les rapports $W_i = Y_i/X_i$. Pour ce qui a trait à notre méthode, si X_i ménages se trouvaient avoir exactement la même composition, nous pourrions les imaginer comme une seule entité dans la liste, avec une fréquence initiale X_i et une fréquence redressée Y_i . Cependant, s'il y a un grand nombre de classes de redressement, il est peu probable que plusieurs ménages du même flot aient exactement la même composition. C'est pourquoi nous ne tenterons pas de grouper les ménages; il sera plus facile, au point de vue de la notation et du calcul, de lister les ménages séparément de sorte que pour la composition de chaque ménage recensé, la fréquence initiale $X_i = 1$ et $Y_i = W_i$. Abstraction faite de la différence de notation, la seule chose qui distingue la formulation mathématique d'un plan de pondération de celle d'un redressement par la méthode itérative du quotient est que les contraintes linéaires n'ont pas la structure particulière des fréquences marginales d'un tableau de contingence. Pour abréger la présentation des exemples, nous indiquerons parfois sur une ligne un nombre qui représentera le nombre de lignes identiques à cette ligne dans la liste des ménages.

Pour ce qui a trait au tableau de contingence, la méthode itérative du quotient préserve les rapports de produits croisés des cases de même que l'indépendance des variables, lorsque celle-ci existe dans le tableau initial. C'est pourquoi, dans le calcul d'estimations régionales, la méthode itérative du quotient est appelée «estimation avec préservation des structures» (Purcell 1979; Purcell et Kish 1979). Voir section 10.1 pour une analyse approfondie des fonctions objectif. Notre méthode diffère de la méthode itérative du quotient en ce que les contraintes linéaires ne se rapportent pas nécessairement aux fréquences marginales d'un tableau de contingence. La méthode itérative du quotient est un cas particulier d'application de notre méthode; il en est de même de la version généralisée de la méthode itérative du quotient de Oh et Scheuren (1978), qui utilise des tableaux différents pour ajuster chaque fréquence marginale. De fait, les covariables continues aussi bien que discrètes peuvent faire l'objet de contraintes; des applications de ce genre sont proposées dans la section 8.3. Par ailleurs, les algorithmes que nous proposons permettent de déterminer directement s'il y a des poids qui sont compatibles avec toutes les contraintes données. Nous pouvons ensuite choisir des contraintes qui doivent être adoucies pour ajuster les poids. Grâce à ces caractéristiques, les possibilités d'application des méthodes proposées vont au-delà de la représentation du sous-dénombrement.

3. AJUSTEMENT DES POIDS

Nous allons maintenant chercher à calculer des poids qui satisfont les contraintes $AW = B$, $W \geq 0$, en minimisant la fonction objectif $T = \sum W_i \log (W_i)$. Pour que T soit une fonction continue de W , nous adoptons la règle usuelle $0 \log 0 = 0$. Nous désignerons tout vecteur de poids qui satisfait les contraintes linéaires (équations et inéquations) comme une *solution réalisable*. Tant que le poids total des ménages est soumis à une contrainte, l'ensemble des solutions réalisables est borné et T suppose par conséquent une valeur minimum pour l'ensemble; de plus, comme T est strictement convexe, la solution est unique. Le calcul des poids se divise donc naturellement en trois étapes: (1) déterminer si les contraintes $AW = B$ sont cohérentes; (2) déterminer s'il y a des solutions réalisables; et (3) trouver la solution réalisable qui minimise T . Nous supposons qu'il y a I ménages et P contraintes de sorte que A est une matrice $P \times I$.

pourquoi on attribue des poids aux *ménages*, les *membres* d'un ménage ayant tous le même poids.

Pour que les chiffres de la liste pondérée correspondent à ceux obtenus par le redressement préalable, les contraintes suivantes doivent être respectées:

- (A1) Dans chaque ilot, la somme des poids des ménages égale le nombre redressé de ménages.
- (A2) Dans chaque classe de redressement et chaque ilot, la somme des poids des personnes égale le nombre redressé de personnes.

Pour que la liste d'ilot pondérée ressemble le plus possible à la liste initiale, il faut aussi respecter la contrainte suivante:

- (B) Les poids doivent être, pour ainsi dire, aussi près que possible les uns des autres.

Avec des poids d'unités (ou poids égaux), la composition de l'ilot demeure inchangée. Si les poids ne sont pas très différents les uns des autres, la composition de l'ilot pour le recensement demeure à peu près la même grâce au plan de pondération. Conformément aux règles d'usage sur l'utilisation des poids dans les enquêtes, nous ne devons pas modifier en profondeur notre représentation de la composition de l'ilot, sauf si les données sur le sous-dénombrement le justifient.

Nous allons maintenant passer à la formulation mathématique de ces critères. Supposons que dans l'ilot considéré, il y a S classes de redressement et I ménages recensés et que le ménage i compte C_{is} membres de la classe s . Supposons que H est le nombre total de ménages que l'on souhaite voir dans la liste redressée pour l'ilot et D_s est le nombre total de personnes que l'on souhaite voir dans la classe s . Soient $W_i, i = 1, 2, \dots, I$, les poids correspondant aux ménages.

(1)
$$\sum_{i=1}^I W_i = H$$

et (A2) exige que

(2)
$$\sum_{i=1}^I W_i C_{is} = D_s, s = 1, 2, \dots, S.$$

Ces contraintes peuvent être représentées par une équation matricielle de la forme $AW = B$,

où

(3)
$$A = \begin{bmatrix} 1 \\ C' \end{bmatrix}, B = \begin{bmatrix} H \\ D \end{bmatrix}, W' = [W_1 W_2 \dots W_I] \text{ et } D' = [D_1 D_2 \dots D_S]$$

et I représente une ligne de uns.

Pour réaliser l'objectif (B), nous choisissons une fonction objectif qui représente la distance entre les poids W et les poids uniformes et nous la minimisons. Nous utiliserons à cette fin la fonction objectif $T = \sum W_i \log (W_i)$. Cette mesure est proportionnelle à l'information discriminante (information de Kullback-Liebler) de la distribution de probabilité discrète (selon les ménages) avec poids relatifs W_i par rapport à la distribution de probabilité avec poids égaux, et elle est la fonction objectif qui est à la base de la méthode itérative du quotient classique (ajustement proportionnel itératif), utilisée pour le redressement des tableaux de contingence (Deming et Stephan 1940; Ireland et Kullback 1968; Oh et Scheuren 1978; ce dernier ouvrage renferme une bibliographie plus complète). Notre méthode peut donc être considérée comme

Extrait d'un fichier de microdonnées d'un échantillon

Tableau 1

Nom	Adresse	Sexe	Âge
John Smith	328 rue Principale	H	34
Mary Smith	328 rue Principale	F	32
Louise Smith	328 rue Principale	F	7
Nancy Chen	330 rue Principale	F	62
Jorge Ramirez	332 rue Principale	H	21
Juan Ramirez	332 rue Principale	H	24

Tableau 2

Fichier de microdonnées reconstruit selon les ménages et montrant la composition des ménages.

Nombre de personnes par classe			
Adresse	Classe 1	Classe 2	Classe 3
328 rue Principale	1	1	1
330 rue Principale	0	1	0
332 rue Principale	2	0	0

Une autre méthode d'imputation consiste à utiliser des modèles probabilistes pour le nombre de personnes non recensées à l'intérieur de ménages et le nombre de ménages non recensés et à calculer un nombre de ménages imputés à partir de la distribution a posteriori des ménages oubliés, étant donné les ménages recensés. Cette méthode est assimilée à l'imputation multiple (Rubin 1987), où tout le processus d'imputation est répété plusieurs fois afin de tenir compte de la variabilité attribuable au sous-dénombrement. Toutefois, dans chaque liste d'ilot ainsi créée, les totaux fondés sur les ménages recensés et imputés ne correspondront pas nécessairement parfaitement aux totaux redressés voulus. Dans cet article, nous nous intéressons surtout à des méthodes qui permettent un ajustement *exact* à des estimations démographiques calculées à une étape antérieure.

Dans les sections qui suivent, nous allons élaborer des méthodes qui permettront d'effectuer le redressement par pondération proposé. La section 2 contient une formulation mathématique des objectifs du plan de pondération tandis que la section 3 expose la façon d'ajuster les poids. Dans la section 4, nous voyons comment faire la distinction dans le plan entre les personnes oubliées à l'intérieur de ménages et les personnes oubliées à cause du sous-dénombrement de ménages. La section 5 renferme quelques perfectionnements qui ont pour but d'accroître la robustesse de la méthode à l'égard de la variabilité des petits îlots. La section 6 contient les résultats de simulations. La section 7 porte sur l'utilisation de données pondérées à diverses fins liées au recensement tandis que la section 8 examine les effets du redressement par pondération sur les covariables qui ne font pas partie du plan utilisé pour la formation des classes de redressement. Enfin, dans la section 9, nous exposons brièvement quelques problèmes non résolus et proposons des sujets de recherche.

2. OBJECTIFS D'UN PLAN DE PONDERATION ET FORMULATION MATHÉMATIQUE

Un des buts essentiels du plan proposé est de répartir la population de l'ilot dans des ménages valides de sorte que les statistiques portant sur les ménages soient clairement définies. C'est

estime de personnes oubliées en proportion du nombre de personnes recensées dans la classe en question. Dans cet article, le mot «redressement» désigne toute opération par laquelle on ajoute le chiffre estimé du sous-dénombrement aux chiffres du recensement. Les classes de redressement peuvent ou non correspondre aux strates formées a posteriori dans l'analyse d'un programme post-censitaire. Ce genre de données pourraient bien se prêter au calcul de fréquences marginales par caractéristique pour les personnes. On pourrait, en particulier, se servir des totaux d'îlots pour calculer les données régionales destinées à divers usages officiels ou commerciaux.

Toutefois, il serait préférable dans certains cas d'inclure les personnes additionnelles dans des ménages. Pour cela, nous supposons qu'il est possible d'estimer le nombre de ménages oubliés dans chaque îlot. Il pourrait aussi y avoir des données qui permettraient de distinguer les personnes qui ont été oubliées à l'intérieur des ménages recensés et les personnes qui font partie des ménages oubliés.

Pour que les chiffres redressés prennent toute leur signification, la composition des ménages ajoutés et les liens entre les membres de chaque ménage doivent être cohérents et représentatifs des caractéristiques des ménages de la région. Le terme «composition» désignera ici le nombre de membres du ménage qui appartiennent à chaque classe de redressement. Par exemple, un ménage formé d'un chef de ménage de sexe féminin et de race blanche âgé de 20 ans, d'une personne de sexe masculin et d'origine chinoise âgée de 75 ans et d'une fille de race noire âgée de 10 ans ne serait pas très vraisemblable même si tous ses membres appartenaient à des classes qui sont bien représentées dans l'îlot. Néanmoins sur le plan abstrait, le fait de définir ces profils et de former des ménages qui leur correspondent est une tâche gigantesque.

Exemple 1: *Constitution d'une liste de ménages.*

Le tableau 1 illustre une portion d'une liste de recensement comme celles pouvant figurer sur une bande de microdonnées.

Le tableau 2 représente la même liste mais cette fois sous forme de résumé de la composition des ménages dans le cas où il existe seulement trois classes d'estimation: (1) hommes de plus de 20 ans, (2) femmes de plus de 20 ans et (3) enfants de 20 ans ou moins.

Essentiellement, le même problème se représente chaque fois qu'il faut répondre les résultats d'une enquête sur les ménages de manière à les rendre compatibles avec les totaux marginaux connus pour diverses catégories de personnes.

L'objet essentiel de la méthode proposée dans cet article est de pondérer les ménages qui figurent sur la liste de recensement de l'îlot de manière que les totaux pondérés de personnes dans chaque classe de redressement et le total pondéré des ménages égalent parfaitement les totaux redressés correspondants. Ainsi, bien que la pondération modifie la composition proportionnelle de l'îlot, tous les ménages sont réels et présentent des caractéristiques et des relations qui sont cohérentes et raisonnables pour l'îlot en question. Cette méthode de pondération ressemble à la méthode itérative du quotient, où le poids appliqué à la fréquence d'une case d'un tableau de contingence est le quotient de la fréquence redressée par la fréquence initiale. Les poids des ménages sont calculés *après* que les totaux d'îlots ont été redressés et sont conformes à ces totaux. Pour la plupart des usages du recensement, les enregistrés pondérés peuvent servir de base à la création de tableaux et de listes.

On peut opposer les méthodes ci-dessus aux méthodes d'imputation, qui consistent à représenter les unités non recensées par des unités entières ajoutées à la liste. Les unités imputées peuvent être des personnes ou des ménages. Bien qu'il soit possible d'imputer des personnes dans l'îlot, on n'a toujours pas trouvé le moyen d'intégrer ces personnes dans des ménages plausibles. L'intégration de ces personnes à des foyers collectifs fictifs, comme cela s'est fait dans certains tests d'opérations de redressement, a pour effet d'étudier le problème en créant une fausse image des relations dans l'îlot. La ré pondération ou l'imputation de personnes seraient des solutions convenables pour les pensionnaires d'institution ou de foyer collectif car dans ce cas, la configuration du ménage n'a pas d'importance.

Redressement des estimations régionales par une pondération des ménages

ALAN M. ZASLAVSKY¹

RÉSUMÉ

Supposons que des taux de sous-dénombrement ont été estimés pour un recensement et que des estimations du niveau de sous-dénombrement ont été établies pour les foyers. Il peut être alors souhaitable de dresser une nouvelle liste de ménages qui comprendrait les ménages qui auraient été oubliés. Nous proposons dans cet article de dresser une telle liste en pondérant les ménages qui représentent le nombre total voulu de personnes dans chaque classe d'estimation et le nombre total voulu de ménages. On calcule alors des poids qui satisfont les contraintes et qui rapprochent le plus possible le tableau des données ajustées des données brutes. On peut voir dans cette méthode un exemple d'application de la méthode itérative du quotient à des cas où les contraintes ne concernent pas les fréquences marginales d'un tableau de contingence. Des covariables continues ou discrètes peuvent être utilisées dans les opérations de redressement et il est possible de vérifier directement si les contraintes peuvent être satisfaites. Enfin, nous posons des méthodes pour l'utilisation de données pondérées à des fins diverses liées au recensement et pour le redressement de données corrélées sur les caractéristiques des ménages oubliés, par exemple le revenu, qui ne sont pas considérées directement dans l'estimation du niveau de sous-dénombrement.

MOTS CLÉS: Sous-dénombrement; méthode itérative du quotient; redressement d'estimations régionales; données manquantes.

1. INTRODUCTION

Beaucoup de recherches ont été faites ces dernières années afin d'élaborer des méthodes pour estimer le taux de sous-dénombrement dans le recensement de 1990 aux États-Unis (National Academy of Sciences 1985). Parmi les principales voies envisagées, notons l'exécution d'une enquête post-censitaire (EP) dans un échantillon d'foyers peu de temps après le recensement. La proportion des personnes enregistrées dans l'EP, qui ne figurerait pas dans les listes du recensement serait une estimation du taux de sous-dénombrement pour le recensement. Des estimations du niveau de sous-dénombrement seraient établies pour un niveau géographique inférieur quelconque (probablement la plus petite unité géographique utilisée pour le recensement, c'est-à-dire l'foyer). Ces estimations se rapporteraient à des classes formées en fonction des caractéristiques des personnes et peut-être de certaines caractéristiques des ménages ou des foyers. Par «classes» nous entendons ici les classes ou cellules d'estimation ou de redressement; par «foyer» nous entendons la plus petite unité géographique pour laquelle il existe des estimations du niveau de sous-dénombrement. Le recensement de 1980 a permis de dénombrer environ 100 millions de ménages dans deux à quatre millions d'foyers, selon les définitions utilisées. Pour chaque foyer, le résultat des méthodes décrites ci-dessus serait un vecteur d'estimations du sous-dénombrement, formé de 5 éléments rattachés à 5 classes de redressement et dont la somme correspondrait au nombre estimé de personnes non recensées pour l'foyer en question. Il n'est pas de notre ressort ici d'exposer les méthodes qui servent à établir ces estimations. Toutefois, dans les exemples que nous utiliserons, nous supposons qu'il existe un taux de sous-dénombrement pour chaque classe dans chaque foyer, lequel taux exprimera le nombre

¹ Alan M. Zaslavsky, Statistics Center, Massachusetts Institute of Technology, Pièce E40-111, Cambridge, MA 02138, E.-U.
E.-U. et Département de statistique, Harvard University, Cambridge, MA 02138, E.-U.

- SEKAR, C.C., et DEMING, W.E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, 44, 101-115.
- SELTZER, W., et ADLAKHA, A. (1974). On the Effect of Errors in the Application of the Chandrasekaran-Deming Techniques. Reprint 14, University of North Carolina, Laboratory for Population Statistics.
- SPENCER, B.D. (1986). Conceptual issues in measuring improvements in population estimates. *Proceedings of the Second Annual Research Conference*, U.S. Bureau of the Census, Washington, D.C., 393-407.
- SPENCER, B.D. (1980). Implications of equity and accuracy for undercount adjustment: a decision-theoretic approach. *Proceedings of the 1980 Conference on Census Undercount*, U.S. Bureau of the Census, Washington, D.C., 204-216.
- WOLTER, K.M. (1986a). Some coverage error models for census data. *Journal of the American Statistical Association*, 81, 338-346.
- WOLTER, K.M. (1986b). A Combined Coverage Error Model for Individuals and Housing Units. SRD Research Report Number Census/SRD/RR-86/27, Statistical Research Division Report Series, U.S. Bureau of the Census, Washington, D.C.

Si $g = 1$, l'estimateur de système dual ne comporte aucun biais car $bgN^{+1}_1/(bN_{11}) = N^{+1}_1/N^{+1}_{11}$.
En ce qui a trait à l'estimation de N^{+1}_1 on peut définir l'erreur due au non-équilibrage par c_p = erreur dans l'estimation du nombre d'enregistrements erronés parce que les secteurs de recherche des échantillons P et D ne coïncident
L'erreur c_p sera non nulle si g n'est pas 1. Le rapport g peut être supérieur ou inférieur à 1. L'erreur est définie $c_p = b(g - 1)N^{+1}_1$.

Mesure

Dans le TOR, c_p a été mesurée lors d'un test qui visait à confirmer que l'équilibrage ne posait pas de problème et que le plan de sondage était sous contrôle. On pouvait dire que le plan de sondage était sous contrôle lorsque le pourcentage d'enregistrements appariés qui se trouvaient dans l'ilot échantillonné était élevé. Comme le plan de sondage était sous contrôle, on peut supposer que g est approximativement égal à 1 et que c_p est négligeable.

Estimation

Le géocodage semble avoir été très bien exécuté pour la région d'essai du TOR. Toutefois, on ne s'est pas attaché à évaluer de façon formelle les effets d'une erreur de géocodage sur l'estimation de EE . Par conséquent, on suppose que g est égal à 1, ce qui implique les relations $E(c_p) = 0$, et $Var(c_p) = 0$.

BIBLIOGRAPHIE

CHILDERS, D., DIFFENDAL, G., HOGAN, H., SCHENKER, N., et WOLTER, K. (1987). The technical feasibility of correcting the 1990 Census. *Proceedings of the Social Statistics Section, American Statistical Association*, 36-45.
CORBY, C., et MULRY, M. (1988). Memorandum à K.M. Wolter, Objet: Matching Error Pilot Study. Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.
COWAN, C.D., et MALEC, D.J. (1986). Capture-recapture models when both sources have clustered observations. *Journal of the American Statistical Association*, 81, 347-353.
DIFFENDAL, G. (1988). Test des opérations de redressement de 1986 dans le Central Los Angeles County *Techniques d'enquête*, 14, 75-92.
ERICKSEN, E.P., et KADANE, J.B. (1985). Estimating the population in a census year: 1980 and beyond. *Journal of the American Statistical Association*, 80, 98-108, 129-131.
HOGAN, H., et MULRY, M. (1987). Operational standards for determining the accuracy of census results. *Proceedings of the Social Statistics Section, American Statistical Association*, 46-55.
HOGAN, H., et WOLTER, K. (1988). Mesure de l'erreur dans une enquête post-censitaire. *Techniques d'enquête*, 14, 105-124.
MORIARTY, C. (1987). STSD 1986 Test Census Memorandum II-12, Objet: Documentation of the Calculation of the Los Angeles Post-Enumeration Survey Block Weights and Dual System Estimate Variances. Statistical Support Division, U.S. Bureau of the Census, Washington, D.C.
SCHENKER, N. (1988). Traitement des données manquantes dans l'estimation de la couverture: le test des opérations de redressement de 1986. *Techniques d'enquête*, 14, 93-104.

l'échantillonnage pour les strates formées a posteriori, la définition d'une distribution pour le paramètre du biais de corrélation et l'établissement de modèles pour l'erreur de déclaration adresse.

REMERCIEMENTS

Les auteurs tiennent à remercier Aref Dajani pour avoir mis au point le logiciel qui a servi aux simulations de même que Chris Moriarty pour avoir calculé les variances. Nous voudrions aussi exprimer notre reconnaissance à Kirk Wolter et Howard Hogan, qui ont contribué à l'élaboration des modèles par leurs conseils et leurs commentaires judicieux. Bruce Spencer remercie également les membres du Undercount Research Staff du U.S. Bureau of the Census pour avoir collaboré à cette étude en vertu d'une Convention sur la statistique (Joint Statistical Agreement).

ANNEXE

Définition de l'erreur d'équilibre

À cause de la non-linéarité de l'estimateur de système dual, il est difficile d'illustrer par un modèle additif les implications techniques de la conciliation des taux de surdénombrement et de sous-dénombrement bruts estimés. C'est pourquoi nous nous appliquons ici à définir un modèle multiplicatif, qui répondra mieux aux exigences de l'estimateur de système dual. Le fait de restreindre les secteurs de recherche pour les échantillons D et P touche deux éléments de l'ESD. D'une part, cette restriction introduit un biais dans l'estimateur du nombre d'enregistrements erronés, *EE*. D'autre part, elle introduit un biais dans l'estimateur du nombre de personnes de l'échantillon P qui ont été recensées, *N*. Pour analyser les effets de cette restriction sur l'estimateur de système dual dans le cadre du Test des opérations de redressement, nous devons définir deux variables:

b = la proportion des enregistrements du recensement exacts qui se trouvent dans le secteur de recherche de l'échantillon P,

g = le rapport entre le nombre d'enregistrements du recensement exacts qui se trouvent dans le secteur de recherche de l'échantillon D et le nombre d'enregistrements exacts qui se trouvent dans le secteur de recherche de l'échantillon P.

La proportion g reflète l'erreur commise dans l'exécution de l'enquête lorsque le secteur de recherche de l'échantillon D diffère de celui de l'échantillon P. Selon la méthode utilisée pour le TOR, $g = 1$. Voyons maintenant ce qui arriverait si g n'était pas égal à 1.

À cause de la limitation du secteur de recherche, seul un pourcentage b des personnes de l'échantillon P qui ont été recensées peuvent être apparées à un enregistrement du recensement. Cela a pour effet d'introduire un biais égal à $(1 - b) N_1$ dans l'estimateur du nombre de personnes de l'échantillon P qui ont été recensées. Par conséquent, le nombre observé de personnes qui figurent à la fois dans l'échantillon P et la liste du recensement est l'estimateur réel de $b N_1$.

De même, la limitation du secteur de recherche ne permet de vérifier l'exactitude que d'un pourcentage b des enregistrements du recensement et de ce nombre, seul un pourcentage g , c'est-à-dire ceux pour lesquels le secteur de recherche coïncide avec le secteur de recherche de l'échantillon D, sera identifié comme correct. Cela a pour conséquence d'introduire un biais égal à $(1 - bg) N_1$ dans l'estimateur du nombre de personnes qui ont été réellement recensées. Ce biais se trouve à l'origine dans l'estimateur du nombre d'enregistrements erronés, *EE*. Par conséquent, le nombre observé de personnes qui ont été réellement recensées est l'estimateur réel de $bg N_1$.

Tableau 10

Distribution a posteriori du taux de sous-dénombrément net selon divers modèles d'imputation acceptables lorsque $\theta = 2.7$

E(S)		Ecart type		B(S)	
TOR		0.23		1.18	
Modèle 000		0.23		0.93	
Modèle 111		0.22		2.79	

On peut estimer la variance totale du taux de sous-dénombrément net estimée en faisant la somme de la variance d'échantillonnage et de la variance non due à l'échantillonnage. Lorsque $\theta = 2.7$, l'écart type observé pour les modèles 000 et 111 (tableau 10) est 0.22, ce qui donne une variance non due à l'échantillonnage de 0.0005 lorsque toutes les erreurs sont prises en considération. L'écart type du taux de sous-dénombrément net estimée est 0.70, ce qui donne une variance d'échantillonnage de 0.49. La variance totale est donc de 0.0054 et l'erreur type de 0.73. Le coefficient de variation du taux de sous-dénombrément net est 0.083. La variance non due à l'échantillonnage a très peu de poids dans la variance totale par rapport à la variance d'échantillonnage.

7. CONCLUSIONS

Lorsqu'on applique la stratification a posteriori dans l'estimation, le taux de sous-dénombrément estimé pour le TOR est de 9.02. La stratification a posteriori a pour effet d'accroître le taux de sous-dénombrément net estimé de 0.6, ce qui est à moins d'un écart type (0.73) de la valeur estimée 8.42. Bien que la stratification a posteriori devrait normalement produire des estimations renfermant un moins haut niveau d'erreur, le résultat est conforme à l'analyse d'erreur.

Puisque nous considérons toutes les sources d'erreur dans la distribution a posteriori du taux de sous-dénombrément net, nous ne connaissons pas la distribution du paramètre du biais de corrélation θ . Nous pourrions supposer une distribution a priori pour θ , mais d'autres pourraient critiquer cette hypothèse. Si nous étions sûrs que θ est égal à 2.7, un intervalle de confiance à 95% pour le taux de sous-dénombrément net serait

$$4.77 < S < 9.55.$$

On obtient cet intervalle en prenant le taux de sous-dénombrément net estimé après stratification a posteriori (9.02) et en le corrigeant en fonction des deux biais estimés 2.79 et 0.93 du tableau 10 et de deux écarts types, 2×0.73 . Nous croyons qu'il s'agit là d'une estimation prudente puisque nous utilisons deux biais estimés différents, issus respectivement des modèles d'imputation 000 et 111. Si nous considérons maintenant toutes les valeurs que peut prendre θ entre 2.1 et 3.7, un intervalle de confiance à 95% pour S serait (4.43, 10.32); ce serait la une estimation très prudente.

Nous croyons que la méthode qui vient d'être exposée dans cet article peut être appliquée dans le recensement de 1990 si on lui apporte les modifications nécessaires. Quant aux recherches futures, elles devraient porter plus spécialement sur l'estimation de l'erreur non due à

Tableau 8
Percentiles de la distribution a posteriori du taux de sous-dénombrement net pour $\theta = 2.7$

	1	5	10	25	50	75	90	95	99
Normale	6.70	6.86	6.94	7.08	7.24	7.40	7.54	7.63	7.79
Uniforme	6.75	6.86	6.93	7.07	7.24	7.42	7.55	7.62	7.73
Gamma	6.67	6.84	6.93	7.08	7.24	7.40	7.53	7.61	7.74

Tableau 9

Distribution a posteriori du taux de sous-dénombrement net pour diverses valeurs de θ

θ	E(S)	Ecart type	B(S)
1.0	5.75	0.18	2.67
2.1	6.72	0.22	1.70
2.7	7.24	0.23	1.18
3.7	8.09	0.27	0.33

Nous avons réalisé des simulations où nous tenions constants les deux premiers moments des erreurs n_p, c_p, m_m, m_f, m_a , et θ mais faisons varier les distributions. Nous avons évalué l'erreur totale lorsque les distributions d'erreur sont toutes des distributions normales, puis toutes des distributions gamma et enfin toutes des distributions uniformes. Le changement de chaque cas, cette distribution se rapprochait sensiblement de la distribution normale. La figure 1 montre la distribution du taux de sous-dénombrement lorsque $\theta = 2.7$, et illustre en même temps les résultats des simulations.

Le tableau 8 donne les percentiles de la distribution du taux de sous-dénombrement net pour divers types de distribution des composantes d'erreur lorsque θ est égal à 2.7 et que la méthode d'imputation du TOR est utilisée. L'écart type de la distribution a posteriori est 0.23. Dans tous les cas, la distribution normale donne une approximation acceptable. L'écart était tout au plus de 0.02 pour les percentiles 5 à 95 et tout au plus de 0.08 pour les percentiles 1 et 99. Le fait de modifier la valeur estimée du paramètre du biais de corrélation θ influe sur les moments de la distribution a posteriori du taux de sous-dénombrement. Cette influence se manifeste dans la moyenne et l'écart type. Le tableau 9 donne les résultats pertinents pour les diverses valeurs de θ , les erreurs étant distribuées suivant une loi normale. Lorsque $\theta = 1$, il n'y a peu près pas de biais de corrélation alors que pour les autres valeurs de θ , nous constatons l'existence de sources d'erreur. Du reste, toutes les sources d'erreur sont prises en considération lorsque $\theta = 2.1, 2.7$, et 3.7. La distribution du taux de sous-dénombrement se déplace vers la droite lorsque la valeur estimée du paramètre du biais de corrélation augmente. On observe aussi une augmentation de la variance. Le biais B(S) est positif pour toutes les valeurs de θ considérées mais diminue à mesure que celles-ci augmentent.

Pour les cas non résolus, les simulations ont été réalisées à l'aide de divers modèles d'imputation acceptables. Malgré quelques variations dans les deux premiers moments de la distribution du taux de sous-dénombrement net, l'estimation du taux de sous-dénombrement net dans le TOR présente une certaine robustesse par rapport aux données manquantes. Le tableau 10 donne les résultats des simulations faites à l'aide des modèles 000 et 111 décrits dans la section 5.7.3. Ces modèles ont permis de définir les limites supérieures et inférieures des estimations du taux de sous-dénombrement suivant tous les modèles d'imputation acceptables. Le biais de l'estimateur du taux de sous-dénombrement net varie de 0.93 à 2.79. Autrement dit, le biais représente entre 11 et 33 % du taux de sous-dénombrement net estimé, qui est de 8.42. Le changement de modèle d'imputation n'a presque pas d'effet sur l'écart type.

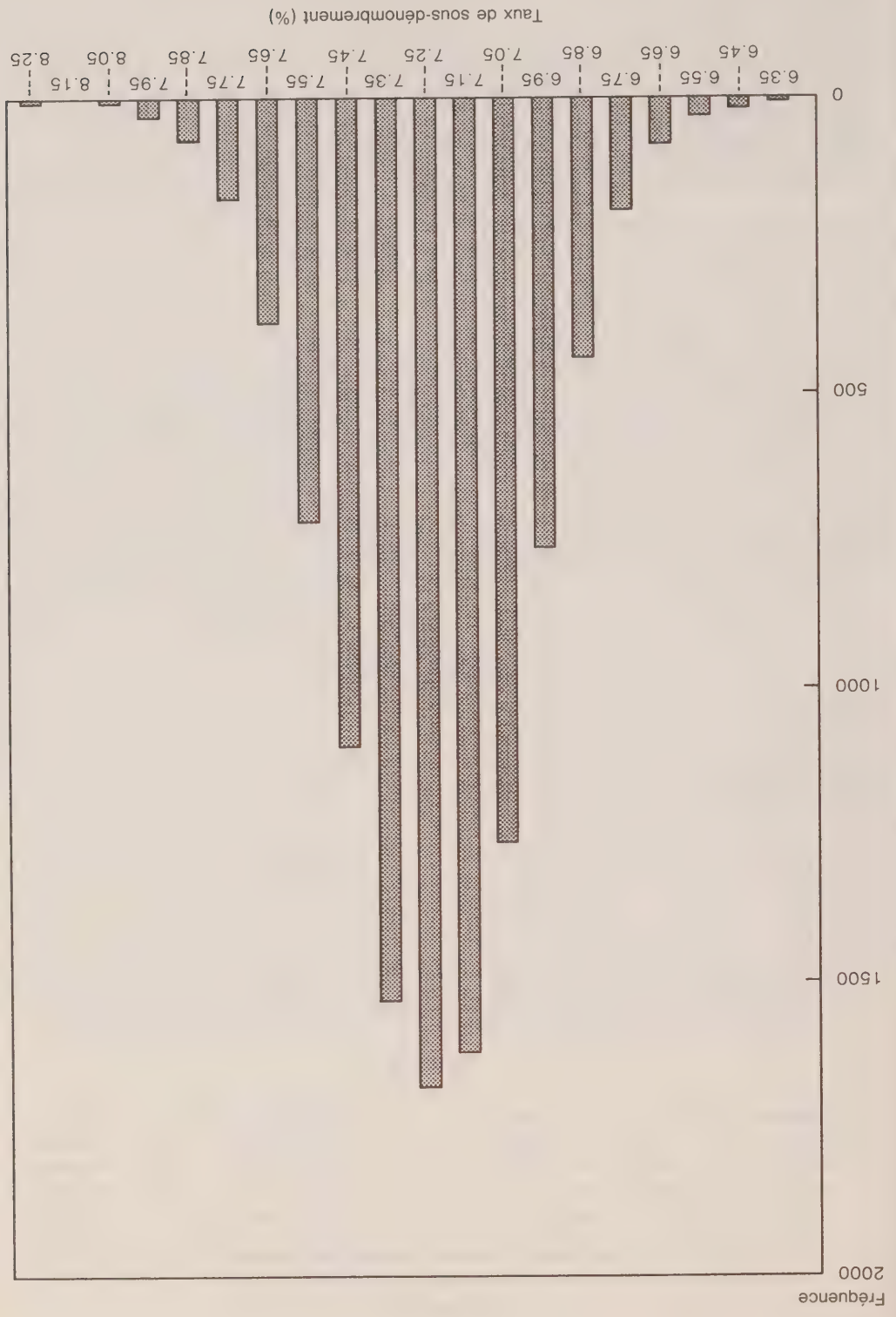


Figure 1. Taux de sous-dénombrement (en pourcentage) lorsque $\theta = 2.7$

Nous avons vu plus haut que le taux de sous-dénombrement net estimé pour la région d'essai du TOR est 8.42 avec une erreur type de 0.7. Nous avons choisi cette estimation car les composantes de l'erreur non due à l'échantillonnage n'ont été estimées que pour la région d'essai en général. Lorsqu'on construit un ESD pour chaque strate formée a posteriori et que l'on additionne ensuite ces estimateurs pour obtenir une estimation à l'échelle de la région, le taux de sous-dénombrement net estimé (en pourcentage) est 9.02.

Le tableau 6.1 donne les moyennes et les écarts types des composantes d'erreur pour l'échantillon de l'EP. Il convient de rappeler que l'ESD pour la région d'essai est 388,040, $M = 298,204$, $C = 343,567$, et $N_p = 336,707$. Nous avons utilisé le poids d'échantillonnage global (17) dans toutes les simulations de sorte qu'il est possible de comparer les effets de diverses hypothèses comme celles touchant les valeurs du paramètre du biais de corrélation, les distributions d'erreur et les modèles d'imputation. La méthode utilisée s'applique aussi dans les cas où l'on utilise un poids d'échantillonnage différent dans chaque strate.

Le tableau 7 indique les effets de chaque catégorie d'erreur sur la distribution a posteriori du taux de sous-dénombrement lorsqu'on applique la méthode d'imputation du TOR. L'erreur d'appariement nette, l'erreur de déclaration d'adresse et l'erreur due à la fabrication sont toutes des erreurs qui influent sur la valeur de M . Par conséquent, la présence d'une seule de ces catégories d'erreur suffit à rendre positif le biais de l'estimation du taux de sous-dénombrement net en pourcentage. L'erreur nette de l'échantillon D est une erreur qui influe sur C. La présence de cette seule catégorie d'erreur suffit à rendre négatif le biais de l'estimation du taux de sous-dénombrement net en pourcentage. Nous avons choisi pour valeur estimée du biais de corrélation la valeur 2.7, qui est la médiane des estimations d'Ericksen et Kadane. La présence du biais de corrélation suffit à rendre négatif le biais de l'estimation du taux de sous-dénombrement net.

Tableau 6

Distributions hypothétiques des estimations d'erreur

Moyenne		Écart type
Écart type		
Erreur d'appariement nette	-1831	176
Erreur de déclaration d'adresse	-3481	510
Erreur due à la fabrication	-2502	244
Erreur nette de l'échantillon D	-238	64

Tableau 7

Effets de chaque catégorie d'erreurs sur la distribution a posteriori du taux de sous-dénombrement net et biais de l'estimateur du taux de sous-dénombrement

E(S)		Ecart type	B(S)
E(S)			
Erreur d'appariement nette	7.86	0.06	0.56
Erreur de déclaration d'adresse	7.35	0.16	1.07
Erreur due à la fabrication	7.34	0.08	1.08
Erreur nette de l'échantillon D	8.49	0.02	-0.07
Biais de corrélation (2.7)	10.61	0.00	-2.19

L'expérience montre que de nombreux secteurs où la population est difficile à dénombrer sont caractérisés par des taux d'omission et d'enregistrement erroné plus élevés.

5.8.2 Mesure

On peut utiliser le modèle de randomisation appliqué couramment aux enquêtes par sondage pour estimer la variance de l'ESD. Le coefficient de variation, qui est le rapport entre la racine carrée de la variance de l'ESD observé et la moyenne de la distribution de l'ESD, renseigne sur l'erreur d'échantillonnage dont est entaché l'ESD.

L'estimateur de la variance de l'estimateur de système dual observé (Moriarty 1987), $v(N^{++})$, est défini (sous forme d'une série de Taylor)

$$v(N^{++}) = N^{++}_2 + v(N^{++}_1)/N^{++}_1 + v(N^{++}_2)/N^{++}_2 - 2c(N^{++}_1, N^{++}_2)/N^{++}_1$$

$$+ N^{++}_2 v(E)/M^2 + 2N^{++}_1 v(E, M)/M^2 - c(E, N^{++}_1)/M,$$

où

$$E = I_2 + EE,$$

$$v(X) = \text{l'estimateur de la variance d'un estimateur } X,$$

$$c(X, Y) = \text{l'estimateur de la covariance de } X \text{ et } Y.$$

Les catérogies I_2 , données insuffisantes pour l'appariement, et EE , enregistrements erronés, sont traitées comme un seul groupe dans l'estimation de la variance. Les estimateurs de la variance et de la covariance reflètent l'échantillonnage d'îlots par grappe et les grappes d'îlots.

5.8.3 Estimation

L'écart type de l'estimation de système dual établie pour la région d'essai (388,040) est 3,100.37. Le coefficient de variation est de 0.008. Par conséquent, l'écart type pour le taux de sous-dénombrement net estimé est 0.7%.

5.8.4 Résumé

L'erreur d'échantillonnage est de 3,100.37 pour l'ESD du TOR et de 0.70% pour l'estimateur du taux de sous-dénombrement net.

6. MODÈLE DE L'ERREUR TOTALE

Nous allons maintenant résumer l'effet combiné des composantes d'erreur à l'aide des distributions *a posteriori* du taux de sous-dénombrement net. On estime le biais du taux de sous-dénombrement net estimé, $B(S)$, en calculant la différence entre et la moyenne de la distribution *a posteriori*. Pour construire une telle distribution, nous avons appliqué une méthode de simulation à 10,000 itérations, qui nous a permis d'obtenir des composantes d'erreur pseudo-aléatoires et de les additionner aux estimations du TOR. En utilisant les formules de la section 5.1.2, nous avons obtenu les équations suivantes:

$$N = (N^{++}_1 - n^{++}_1) + (C + c - (M - m))$$

$$+ \theta(C - c - (M - m))(N^{++}_1 - n^{++}_1) + (M - m)/(M - m)$$

$$= (C - c - (M - m))(N^{++}_1 - n^{++}_1) + (M - m)/(M - m)$$

$$+ (\theta - 1)(C - c - (M - m))(N^{++}_1 - n^{++}_1) + (M - m)/(M - m).$$

Plusieurs distributions ont été utilisées pour reproduire diverses estimations de l'erreur d'imputation et du biais de corrélation (θ pour paramètre) et diverses distributions marginales pour les composantes - à savoir normale, gamma et uniforme.

indique dans quelle mesure l'ESD est sensible à la méthode d'imputation. Par exemple si l'intervalle est étroit, on en déduit que les estimations sont robustes et que les données manquantes n'en modifient pas réellement la fiabilité.

5.7.3 Estimation

On a évalué l'effet des données manquantes sur les estimations du TOR de 1986 en étudiant l'intervalle des estimations obtenues lorsqu'on utilise, à la place de la méthode désignée, des méthodes d'imputation fondées sur d'autres hypothèses plausibles. Nous parlons ici plus précisément de diverses méthodes de traitement des cas d'interview par personne interposée, des cas de déménagement et des cas désignés comme enregistrés fictifs (Schenker 1988). Suivant ces méthodes, on a classé les cas d'interview par personne interposée de l'échantillon P dans les cas de non-interview et on a fait un redressement par pondération. Ainsi, les cas d'interview par personne interposée avaient essentiellement le même taux d'appariement que les cas d'interview sans personne interposée. En ce qui a trait aux cas de déménagement de l'échantillon P, on les a tous classés dans les cas non résolus puis on a imputé des probabilités d'appariement pour tous ceux-là au lieu d'en imputer uniquement pour les cas réellement non résolus. Ainsi, les personnes ayant déménagé avaient essentiellement le même taux d'appariement que les personnes n'ayant pas déménagé. En ce qui a trait aux enrégistrement fictifs, ils ont été identifiés par suite d'un examen des cas non résolus de l'échantillon D confié à des spécialistes de l'appariement, qui ont reconnu des cas non résolus comme des cas fictifs. Cela a eu pour conséquence de hausser les taux observé et imputé d'enrégistrement erroné. Les modèles 000 et 111 du tableau 4 de l'article de Schenker donnent, respectivement, les limites supérieures et inférieures des taux de sous-dénombrement estimés. Les deux modèles diffèrent du TOR en ce que les migrants internes tiennent la place des migrants externes. Les migrants internes de l'échantillon P sont les répondants de cet échantillon qui ont déménagé entre le jour du recensement et le jour de l'interview de l'EP. Dans le TOR de 1986, les migrants internes de l'échantillon P qui venaient de l'extérieur de la région d'essai n'ont pas été inclus dans l'estimation de l'EP. Le fait d'exclure les migrants externes de l'estimation revient essentiellement à supposer qu'ils avaient le même taux de saisie dans le recensement que les personnes incluses dans l'estimation. Habituellement, les personnes ayant déménagé ont un taux de saisie moindre que celles n'ayant pas déménagé. Le modèle 000 correspond à la méthode utilisée dans le TOR tandis que le modèle 111 correspond à toutes les autres méthodes.

5.7.4 Résumé

On évalue l'effet des données manquantes sur la distribution de l'erreur totale en calculant les paramètres de la distribution du taux de sous-dénombrement selon plusieurs modèles d'imputation acceptables. Les modèles qui déterminent les limites supérieure et inférieure du taux de sous-dénombrement servent à l'analyse de l'erreur totale.

5.8 Erreur d'échantillonnage

5.8.1 Source d'erreur

L'ESD observé est exposé à l'erreur d'échantillonnage car N_p , C et M sont déterminés à partir d'échantillons. La taille de l'échantillon de l'EP dépend du degré d'erreur d'échantillonnage toléré et des sommes allouées pour l'enquête. Toutes choses étant égales par ailleurs, plus l'échantillon sera grand, moins il y aura d'erreur d'échantillonnage dans les estimations. L'estimateur et le plan d'échantillonnage influent sur l'erreur d'échantillonnage. Selon le plan de l'EP réalisée dans le cadre du TOR, les données des échantillons P et D sont tirées du même échantillon d'îlots. L'échantillon P comprend toutes les personnes qui habitent dans les unités de logement contenues dans les îlots échantillonnés. L'échantillon D comprend tous les enrégistrement du recensement qui se rapportent aux îlots échantillonnés. L'estimation de l'erreur d'échantillonnage tient compte de la corrélation qui tend à exister entre les cas d'omission et les cas d'enrégistrement erroné à l'intérieur des îlots et des unités de logement.

L'intermédiaire d'enregistrements erronés. L'échantillon P permet de mesurer le taux de sous-dénombrement brut par l'intermédiaire des personnes qui n'ont pas été recensées. Idéalement, il faudrait examiner tous les enregistrements du recensement avant de dire qu'une personne de l'échantillon P n'a pas été recensée. De même, il faudrait parcourir la liste de tous les habitants du pays avant de déterminer si un enregistrement de l'échantillon D est un enregistrement répété ou fictif. Evidemment, de telles recherches sont tout simplement impossibles dans le cadre d'une EP. Il est donc nécessaire de les circonscrire dans des limites raisonnables. Il faut du même coup conserver l'erreur nette bien qu'il faille s'attendre à une hausse des taux de surdénombrement et de sous-dénombrement bruts mesurés à cause de la limitation du secteur de recherche. Les taux de surdénombrement et de sous-dénombrement bruts doivent s'équilibrer pour évaluer l'erreur de couverture nette.

Le fait de ne pas réaliser l'équilibre entre le taux de surdénombrement brut estimé et le taux de sous-dénombrement brut estimé peut fausser le nombre d'enregistrements de l'échantillon D qui sont classés par erreur dans les enregistrements erronés. Ce genre d'erreur peut créer soit un biais par excès ou un biais par défaut.

Contrairement à ce que ce fut en 1980, la notion d'équilibrage ne figure pas dans le plan de sondage de l'EP prévue pour 1990 et testée dans le TOR de 1986. Ce plan de sondage prévoit le chevauchement des échantillons P et D. Les mêmes ilots sont contenus dans les deux échantillons. Le secteur de recherche pour l'échantillon P est défini comme le secteur de recherche fondamental. On choisit le secteur de recherche pour l'échantillon D de manière à ce qu'il soit conforme au secteur de recherche pour l'échantillon P.

5.6.2 Résumé

L'erreur de géocodage dans le TOR de 1986 est jugée négligeable et n'est donc pas incluse dans le modèle de l'erreur totale. On trouvera en annexe un modèle de l'erreur d'équilibrage.

5.7 Données manquantes

5.7.1 Source d'erreur

Les échantillons P et D ne sont pas à l'abri des données manquantes. Il y a des cas dans l'échantillon D où l'on ne dispose pas des données nécessaires pour déterminer si la personne en question a été enregistrée correctement au cours du recensement. De même, il y a des cas dans l'échantillon P où, à cause d'un manque de données, on ne peut dire si la personne a été réellement recensée. Devant l'incapacité de résoudre le cas, on impute statistiquement la probabilité que la personne soit recensée.

Plusieurs situations peuvent faire qu'un code de dénombrement est indéterminé. L'interviewer peut ne pas être en mesure de réaliser l'interview avec une personne de l'échantillon P ou l'interview de rappel. Un questionnaire de l'échantillon P ou de l'échantillon D peut ne pas contenir toutes les données démographiques et les données sur le logement nécessaires à l'estimation. Même si des questionnaires renferment tous les renseignements voulus, les circonstances peuvent être à ce point obscures qu'il n'est pas possible de déterminer un code de dénombrement.

5.7.2 Mesure

Au lieu de considérer séparément les composantes c_i , m_i et n_{pi} , nous évaluons l'erreur de l'ESD attribuable aux données manquantes. Notre méthode consiste à soumettre divers modèles de compensation de la non-réponse jugés acceptables à une analyse de sensibilité. Tout d'abord, avant même l'exécution de l'EP, nous choisissons une méthode jugée préférable d'imputation pour les cas non résolus des échantillons P et D. Selon les problèmes qui surgissent durant la collecte et le traitement des données de l'EP, nous pouvons proposer diverses méthodes de traitement des données manquantes jugées acceptables. On peut alors calculer l'ESD suivant ces divers modèles de compensation de la non-réponse. L'intervalle des estimations possibles

5.5.2 Définition

Le biais de l'ESD attribuable à des codes de dénombrement erronés découle de l'erreur d'estimation de N_{+1} . Lorsqu'on estime le nombre de personnes qui ont été réellement recensées C , on applique un facteur de correction pour le nombre d'enregistrements erronés EE , EE et, par conséquent, C sont calculés à l'aide des données de l'échantillon D . On commet une erreur d'estimation pour C lorsqu'on attribue le mauvais code de dénombrement à des cas de l'échantillon D . Soit

c_e = la différence entre le nombre pondéré d'enregistrements erronés considérés par erreur comme exacts et le nombre pondéré d'enregistrements exacts considérés par erreur comme erronés.

L'espérance de c_e , étant donné la valeur observée C , est désignée par $E(c_e)$. La variance de c_e , étant donné la valeur observée C , est désignée par $Var(c_e)$.

5.5.3 Mesure

On peut évaluer directement l'erreur administrative en soumettant un échantillon de cas à un nouvel appariement. Les autres types d'erreur, comme la répétition d'enregistrements qui découle de la violation des règles concernant le lieu du domicile le jour du recensement, peuvent être évalués par un examen des distributions de fréquence des enregistrements erronés. Cette méthode est préférable à une méthode d'évaluation directe à cause de la difficulté à obtenir des données précises dans les opérations de suivi subséquentes. Lorsque des tests confirment que l'erreur brute correspondante est à un niveau acceptable, on peut supposer que l'erreur nette est négligeable. Par exemple, si nous répartissons les enregistrements erronés selon le groupe d'âge, nous devrions normalement observer un grand nombre d'enregistrements répétés au sein des groupes de la population qui sont très mobiles et pour lesquels il y a plus de chances de ne pas respecter les règles concernant l'adresse du domicile le jour du recensement. Dans le TOR de 1986, on a évalué simultanément les opérations de traitement de l'échantillon D et les opérations d'appariement de l'échantillon F dont il a été question dans la section 5.2.3 (Corby et Mulry 1988). Les données pour l'échantillon D , tirées du même sous-échantillon de 35 îlots, ont été traitées à nouveau.

5.5.4 Estimation

Nous allons maintenant estimer les moments de la distribution de c_e d'après l'échantillon de l'EP. Les résultats du second traitement (Hogan et Wolter 1988) indiquent un taux d'erreur nette de 0.0007 en ce qui concerne l'identification des enregistrements exacts. L'espérance de c_e est $E(c_e) = -238$. Cette estimation n'est fondée que sur les cas résolus de l'échantillon D car l'erreur d'imputation pour les cas non résolus est étudiée dans la section 5.7 - Données manquantes.

On n'a pas calculé d'estimation de la variance de l'erreur nette. À notre avis, il est possible d'établir une estimation prudente de la variance en appliquant les définitions de la section 5.2.2. Par conséquent, la variance estimée pour la région d'essai est $Var(c_e) = (17)^2 \times 14 = 4,046$. Pour le modèle de l'erreur totale, les deux premiers moments de la distribution a posteriori de l'erreur nette dans l'estimation du nombre d'enregistrements exacts sont supposés être $E(c_e) = -238$ et $Var(c_e) = 4,046$.

5.5.5 Résumé

5.6 Conciliation des estimations des taux de surdénombrement et de sous-dénombrement bruts

5.6.1 Source d'erreur

L'échantillon D et l'échantillon F servent tous deux à évaluer les erreurs d'enregistrement dans le recensement. L'échantillon D permet de mesurer le taux de surdénombrement brut par

directement r_{fm} , nous supposons avec une certaine prudence qu'il existe une similitude entre les personnes qui ont été réellement interviewées et celles qui ne l'ont pas été mais qui auraient dû l'être. C'est pourquoi nous disons que r_{fm} est égal au taux d'appartenance global pour l'échantillon P.

Les résultats du suivi exécuté après la production des estimations initiales nous permettent de conclure à un taux de fabrication d'environ 1.2%. Le taux d'appartenance pour le TOR est de 88.6% (Diffendal 1988). Par conséquent, l'espérance de l'erreur m_f est $E(m_f) = -2502$. On n'a pas calculé d'estimation de la variance de l'estimateur de l'erreur due à la fabrication. À notre avis, il est possible d'obtenir une estimation prudente de la variance en appliquant les définitions de la section 5.4.2. Ainsi, la variance estimée pour la région d'essai est

$$\text{Var}(m_f) = (17)^2 \times 206 = 59,534.$$

5.4.5 Résumé

Pour le modèle de l'erreur totale, les deux premiers moments de la distribution de l'erreur nette due à la fabrication d'interviews sont supposés être $E(m_f) = -2502$ et $\text{Var}(m_f) = 59,534$. L'erreur nette due à la fabrication de ménages dans l'échantillon P est supposée négligeable et, par conséquent, $E(n_{pf}) = 0$ et $\text{Var}(n_{pf}) = 0$.

5.5 Estimation du nombre d'enregistrements erronés

5.5.1 Source d'erreur

Des enregistrements peuvent se trouver parmi les enregistrements du recensement à la suite d'erreurs. C'est ce qu'on appelle des enregistrements erronés. Comme il faut estimer le nombre de personnes réellement recensées pour calculer l'estimateur de système dual, l'estimation de la population totale comporte un facteur de correction pour les enregistrements erronés. En soustrayant des chiffres du recensement le nombre estimé d'enregistrements qui ne se rapportent pas à une seule personne, on obtient une meilleure estimation du nombre de personnes réellement recensées. Cette correction est estimée au moyen de l'échantillon D dans l'EP.

Les enregistrements erronés comprennent les catégories suivantes: (1) personnes décédées avant le jour du recensement, (2) personnes nées après le jour du recensement, (3) enregistrements qui ne se rapportent pas à des personnes réelles, (4) enregistrements répétés, (5) personnes qui ont été recensées dans un autre secteur que celui où est tenté l'appariement. Le secteur de recherche comprend habituellement l'ilot où est située l'adresse en question de même que l'anneau d'ilot avoisinants.

Le type d'erreur étudiée dans la présente section survient lorsqu'on estime l'erreur de couverture dans le recensement. On commet une erreur dans l'estimation du nombre d'enregistrements erronés lorsqu'on considère un enregistrement de l'échantillon D comme erroné alors qu'il est exact ou vice-versa. Il peut donc se produire des erreurs dans les deux sens lorsqu'on estime le nombre d'enregistrements répétés et les enregistrements fabriqués sont les deux catégories d'enregistrements erronés qui ont le plus de chances de ne pas être reconnus comme tels. Nous ne considérons ici que les erreurs d'estimation se rapportant à ces deux catégories car les erreurs qui touchent les autres catégories sont ou bien sans importance ou bien considérées dans une autre section. Les erreurs d'estimation touchant les enregistrements qui concernent des personnes décédées avant le jour du recensement ou des personnes nées après ce jour ont un effet dérisoire. L'erreur d'estimation touchant les personnes qui ont été recensées dans un autre secteur que le secteur de recherche est traitée dans la section 5.6 – Conciliation des estimations des taux de surdénombrement et de sous-dénombrement bruts.

5.3.5 Résumé

Pour le modèle de l'erreur totale, les deux premiers moments de la distribution de l'erreur due à une mauvaise indication de l'adresse au jour du recensement pour l'échantillon de l'EP sont supposés être $E(m_a) = -3481$ and $\text{Var}(m_a) = 260,100$.

5.4 Fabrication dans l'échantillon P

5.4.1 Source d'erreur

Les intervieweurs peuvent parfois inventer des personnes dans des unités de logement de l'échantillon P. Des études ont montré que la fabrication dans l'interview de l'EP peut introduire un biais appréciable dans les estimations de l'erreur de couverture du recensement fondées sur l'estimateur de système dual. Essentiellement, la fabrication d'enregistrements peut avoir pour effet de réduire le taux d'appariement pour l'EP, contribuant par le fait même à gonfler l'estimation de l'erreur de couverture.

Au U.S. Bureau of the Census l'expérience montre que c'est la fabrication de ménages entiers qui pose le plus de problèmes dans les enquêtes-ménages. Il est rare que des interviewers ajoutent une personne fictive dans un ménage qui existe réellement. Le contrôle qualitatif portant sur l'interview de l'échantillon P vise à découvrir les interviews fictives et à interviewer les vrais membres du ménage. Par conséquent, l'établissement des estimations de système dual ne comporte aucune correction statistique pour la fabrication dans l'échantillon P.

5.4.2 Définition

Les éléments N_{11} et N_{1+} de l'estimateur de système dual sont estimés à l'aide des données de l'échantillon P. Nous avons vu dans la section 4 que:

$$m_f = \text{le nombre pondéré de personnes qui ont été recensées mais auxquelles on a substitué des enregistrements fabriqués};$$

$$n_{pf} = \text{l'erreur dans } N_{pf} \text{ due à la fabrication de ménages dans l'échantillon P.}$$

Les espérances et variances a posteriori de m_f et n_{pf} sont désignées par $E(m_f)$ et $E(n_{pf})$ et $\text{Var}(m_f)$ et $\text{Var}(n_{pf})$ respectivement.

5.4.3 Mesure

Dans le TOR de 1986, le contrôle qualitatif des interviews a permis d'établir un taux de fabrication d'environ 0.6%. Un suivi exécuté après la production des estimations initiales a permis d'établir un taux de fabrication approximatif de 1.2% (Hogan et Wolter 1988).

5.4.4 Estimation

Nous allons maintenant estimer les moments des distributions a posteriori de n_{pf} et de m_f tirées de l'échantillon de l'EP. Nous jugeons raisonnable de supposer que la valeur de n_{pf} dans le TOR est négligeable. Par conséquent, l'espérance et la variance de n_{pf} sont définies $E(n_{pf}) = 0$ et $\text{Var}(n_{pf}) = 0$.

Les données du contrôle qualitatif peuvent servir à estimer $r_f = \text{le taux de fabrication des interviews dans l'échantillon P.}$

En cherchant les enregistrements du recensement qui se rapportent aux personnes de l'échantillon P qui, selon le contrôle qualitatif, n'ont pas été interviewées alors qu'elles auraient dû l'être, on obtient

$$r_{fm} = \text{le taux d'appariement pour les personnes qui n'ont pas été interviewées parce que leur ménage a fait l'objet de fabrication dans l'échantillon P.}$$

Comme rien n'a été consigné dans le TOR, on ne peut connaître l'identité des personnes qui n'ont pas été interviewées telles que repérées au cours du contrôle qualitatif. En conséquence, aucun appariement n'a été tenté. Puisqu'il n'existe pas de données permettant d'estimer

5.3.2 Définition

Le dénominateur N_{11} de l'estimateur de système dual est estimé à l'aide des données de l'échantillon P. Nous avons vu dans la section 4 que :

A_{31} = le nombre pondéré de personnes qui ont été recensées mais dont l'adresse au jour du recensement est inexacte ;
 B_{31} = le nombre estimé de personnes dont l'adresse au jour du recensement est inexacte mais dont l'enregistrement a été repéré à une autre adresse.

Par conséquent, l'erreur nette due à l'inexactitude de l'adresse au jour du recensement, m_a , peut être définie $m_a = B_{31} - A_{31}$. L'espérance et la variance de m_a , étant donné la valeur observée N , sont désignées par $E(m_a)$ et $\text{Var}(m_a)$, respectivement.

5.3.3 Mesure

L'évaluation de m_a repose sur un suivi effectué auprès d'un sous-échantillon de répondants de l'échantillon P qui sont classés parmi les personnes non recensées. Les résultats du suivi servent à estimer l'erreur qui se produit lorsque des personnes qui ont été recensées indiquent mal l'adresse où elles demeureraient le jour du recensement lorsqu'elles répondent à l'EP. On a évalué l'importance de l'erreur de déclaration d'adresse à la suite du TOR de 1986. Après la production des estimations initiales, on a réalisé une interview de rappel auprès d'un échantillon de 903 personnes qui n'avaient pu être apparées afin de déterminer le nombre de cas de non-concordance attribuables à une erreur de déclaration d'adresse. En ce qui concerne les répondants qui ont dit avoir déménagé sans quitter la région d'essai, on a tenté un apparement à la nouvelle adresse.

5.3.4 Estimation

Les cas échantillonnés pour lesquels on a constaté une erreur de déclaration d'adresse peu-vent servir à estimer L_e = le nombre pondéré de personnes qui, durant l'interview de l'échantillon P, ont mal indiqué l'adresse où elles demeureraient le jour du recensement. Une recherche d'enregistrements du recensement aux nouvelles adresses produit r_{am} = l'estimateur du pourcentage de personnes qui ont mal indiqué l'adresse où elles demeureraient le jour du recensement mais qui peuvent être apparées à des enregistrements du recensement.

Par conséquent, l'espérance de l'erreur m_a est estimée par

$$E(m_a) = -r_{am}L_e$$

Selon les résultats de l'interview de rappel (Hogan et Wolter 1988), il y aurait eu erreur de déclaration d'adresse dans tout au plus 3.1% des cas de l'échantillon P. On a calculé un taux d'appariement estimé de 33% pour les personnes qui ont mal indiqué l'adresse où elles demeureraient le jour du recensement et qui ont déménagé sans quitter la région d'essai. Si nous supposons un taux d'appariement identique pour les personnes qui ont dit avoir demeuré à l'extérieur de la région d'essai le jour du recensement, l'espérance mathématique est donc $E(m_a) = -3481$. On n'a pas estimé la variance de l'erreur due à une mauvaise indication de l'adresse. À notre avis, 900 serait une estimation prudente de la variance pour l'échantillon de l'EP. Par conséquent, la variance pour la région d'essai est

$$\text{Var}(m_a) = (17)^2 \times 900 = 260,100.$$

5.2.4 Estimation

Nous allons maintenant utiliser les résultats des sous-échantillons qui ont servi aux études précitées pour estimer les moments de la distribution de m^m tirée de l'échantillon de l'EP. Le fait de ne pas effectuer de recherche approfondie dans l'étude portant sur les personnes n'ayant pas déménagé a probablement contribué à sous-estimer le nombre de cas de fausse non-concordance. Les résultats de recherches approfondies antérieures nous indiquent qu'une façon modérée de pallier à l'absence de telles recherches est d'augmenter de 20% l'erreur nette, qui est de 70 en l'occurrence (Hogan et Wolter 1988). D'après les résultats des deux études, l'erreur nette dans l'échantillon de l'EP est de 95. Le taux d'erreur nette est donc $-0,0055$. Nous appliquons ce taux uniquement aux cas de l'échantillon F qui ont été résolus car l'erreur d'imputation pour les cas non résolus est traitée dans la section 5.7 - Données manquantes. L'espérance de m^m devient $E(m^m) = -1831$, lorsque le poids d'échantillonnage global (17) est utilisé. On n'a pas calculé d'estimation de la variance de l'estimateur de l'erreur d'appariement nette pour les personnes n'ayant pas déménagé. La variance d'échantillon du nombre d'erreurs pour les personnes ayant déménagé est nulle car toutes les personnes de cette catégorie qui n'avaient pu être appariées ont fait l'objet d'un nouvel appariement. Cependant, nous ne croyons pas que la variance réelle est nulle. On pourrait définir une variance en supposant que les erreurs se sont produites suivant une combinaison de processus de Poisson, par exemple les erreurs d'appariement pour les personnes ayant déménagé ayant suivi un processus de Poisson particulier et les erreurs d'appariement pour les personnes n'ayant pas déménagé ayant suivi indépendamment un autre processus de Poisson. Si on considérait que toutes les erreurs suivent un seul et même processus de Poisson, on obtiendrait en définitive une estimation prudente de la variance; en l'occurrence, cette estimation serait le produit de 17×107 . Cependant, le modèle de Poisson ne produit pas nécessairement d'estimation prudente si les erreurs se produisent dans des grappes. Dans une tentative pour établir des estimations prudentes de la variance, nous avons multiplié (quelque peu arbitrairement) l'estimation de la variance calculée suivant un seul processus de Poisson par le poids d'échantillonnage global, ce qui a donné

$$\text{Var}(m^m) = (17)^2 \times 107 = 30,923.$$

5.2.5 Résumé

Pour le modèle de l'erreur totale, les deux premiers moments de la distribution a posteriori de l'erreur d'appariement nette pour l'échantillon de l'EP sont supposés être $E(m^m) = -1831$ et $\text{Var}(m^m) = 30,923$.

5.3 Erreur de déclaration de l'adresse au jour du recensement

5.3.1 Source d'erreur

Quelques-uns des répondants de l'échantillon F ont déménagé entre le jour du recensement et le jour de l'interview de l'EP. Les réponses fournies à cet égard peuvent être erronées. Par exemple, si le répondant a déménagé, l'adresse antérieure qu'il a indiquée peut être inexacte ou peut avoir été mal géocodée par les préposés du codage. L'une ou l'autre de ces erreurs peut fausser le processus d'appariement en orientant la recherche d'enregistrements dans une région différente de celle où le répondant a été recensé. Ainsi, des répondants qui ont été effectivement recensés pourraient être considérés comme des cas de non-concordance parce que les opérations d'appariement ne permettent pas de repérer les enregistrements correspondants. En classant par erreur des répondants parmi les cas de non-concordance, on introduit un biais par excès dans l'estimation du nombre de personnes oubliées dans le recensement. Il arrive qu'une erreur de déclaration d'adresse ne se traduise pas par une fausse non-concordance. Si l'adresse au jour du recensement est dans le secteur où doit normalement se trouver l'adresse déclarée et que celle-ci est géocodée correctement, il y aura appariement.

5.2 Erreur d'appariement

5.2.1 Source d'erreur

Dans la présente analyse, l'erreur d'appariement désigne les erreurs qui sont commises lorsqu'on compare les enregistrements de l'échantillon P à ceux du recensement. L'erreur d'appariement n'englobe donc pas les erreurs de réponse qui surviennent durant la collecte des données. Bien que d'autres types d'erreur puissent entraîner l'attribution d'un mauvais code de dénombrement à un répondant de l'échantillon P, ces sources sont traitées sous d'autres rubriques.

Une fois l'interview de l'échantillon P terminée, on parcourt les enregistrements du recensement pour vérifier si les répondants en question ont bel et bien été recensés. On attribue alors un code à chacun des répondants selon qu'il a pu être apparié ou non à un enregistrement du recensement. Lorsqu'on attribue le mauvais code de dénombrement à une personne de l'échantillon P durant le traitement des données, on commet ce qu'on appelle une erreur d'appariement. Il y a deux façons de commettre des erreurs d'appariement. Soit qu'une personne est appariée à un enregistrement du recensement alors qu'elle n'a pas été recensée à l'origine (c'est ce qu'on appelle une «fausse concordance»), soit qu'elle est reconnue comme non recensée alors qu'elle l'a effectivement été (c'est ce qu'on appelle une «fausse non-concordance»). L'erreur d'appariement aura pour effet de fausser l'estimation de l'effectif recensé et de l'effectif de l'échantillon P et introduira par conséquent un biais dans l'estimation du nombre de personnes oubliées dans le recensement.

5.2.2 Définition

Le dénominateur N_{11} de l'estimateur de système dual est estimé à l'aide des données de l'échantillon P. Nous avons vu dans la section 4 que:

A_{21} = le nombre pondéré de personnes qui ont été recensées,

B_{21} = le nombre estimé de personnes qui ont pu être appariées.

Par conséquent, l'erreur nette due à l'attribution du mauvais code de dénombrement, m_m , peut être définie $m_m = B_{21} - A_{21}$. L'espérance mathématique et la variance de m_m étant donné la valeur observée \hat{M} sont désignées par $E(m_m)$ et $Var(m_m)$.

5.2.3 Mesure

On peut mesurer m_m en traitant une seconde fois une partie de l'échantillon P, c.-à-d. en confiant à du personnel hautement qualifié la tâche d'exécuter un nouvel appariement. On suppose pour cela que des personnes mieux entraînées commettent moins d'erreur même si elles disposent des mêmes documents et des mêmes données qui ont servi à l'appariement initial. Il est possible de rapprocher les nouveaux codes d'appariement et les codes d'appariement initiaux et d'éliminer les différences.

Deux études du processus d'appariement ont été réalisées à l'aide des données du TOR de 1986. Ces études ont servi à évaluer l'appariement pour les personnes ayant déménagé et les personnes n'ayant pas déménagé respectivement.

Dans la seconde étude (Corby et Mulry 1988), on a prélevé un sous-échantillon probabiliste de 35 flots pour le soumettre à un nouvel appariement; cette opération a été confiée à des spécialistes du bureau central. L'échantillon ainsi obtenu a été stratifié selon le taux d'appariement et des taux de sondage exceptionnellement élevés ont été appliqués aux flots qui présentaient de faibles taux d'appariement de telle manière que les personnes affectées au contrôle de la qualité pouvaient recueillir le plus de renseignements possible sur les erreurs d'appariement. Comme les flots adjacents n'ont pas été analysés, il se peut que l'on ait sous-estimé le nombre de cas de fausse non-concordance.

L'autre étude a servi à évaluer l'erreur d'appariement pour les personnes ayant déménagé (Childers et coll. 1987). Quatre-vingt-dix personnes ayant déménagé n'ont pu être appariées dans le TOR et ces 90 cas ont été soumis à un nouvel appariement. Cette opération a permis de découvrir onze cas de concordance additionnels, dont deux n'avaient pas été retenus à la suite du contrôle informatique.

La formulation de Wolter prévoit une action individuelle (autonomie) ou collective (absence d'autonomie) des membres des ménages. Pour leur part, Cowan et Malec présentent un modèle qui permet de grouper les cas d'omission dans le recensement (absence d'autonomie). Nous allons maintenant décrire par un modèle l'effet combiné des sources du biais de corrélation sur l'ESD.

5.1.2 Définition

Pour mieux connaître l'effet du biais de corrélation, supposons que $\theta_i = \theta$ pour tous les i et exprimons la taille réelle de la population par la formule

$$N = N_{11} + N_{12} + N_{21} + \theta (N_{12}N_{21}/N_{11}),$$

où θ_i est le rapport des produits croisés défini dans la section 5.1.1.

Le biais de corrélation influe uniquement sur le dernier terme de l'expression car les trois autres peuvent être estimés directement. Le paramètre θ , représente l'effet de la non-vérification des hypothèses d'indépendance. Lorsque celles-ci se vérifient, $\theta = 1$.

Le biais de corrélation, qui existe lorsque θ est différent de 1, est la seule composante de t , qui est l'erreur due à la non-applicabilité du modèle. La taille de la population peut être définie par la formule:

$$N = N_{11} + N_{+1}/N_{11} + t = N_{11} + N_{+1}/N_{11} + (\theta - 1)(N_{12}N_{21}/N_{11}).$$

En conséquence, le biais de corrélation $t = (\theta - 1)(N_{12}N_{21}/N_{11})$.

5.1.3 Mesure

On peut estimer le paramètre θ au niveau national pour des sous-groupes raciaux et ethniques en utilisant des estimations de la population tirées d'une analyse démographique. L'utilisation de cette méthode suppose toutefois qu'il s'agit là d'estimations précises. Malgré cela, cette formulation permet aussi de faire varier θ afin d'évaluer la sensibilité de l'ESD à l'effet estimé du non-respect des hypothèses d'indépendance.

5.1.4 Estimation

On n'a pas calculé d'estimation de θ pour le TOR de 1986 parce qu'il n'y avait pas d'autre source d'estimations démographiques (par ex., on ne pouvait tirer d'estimations d'une analyse démographique). Cependant, Erickson et Kadane (1985) ont calculé trois estimations de θ pour la population noire en ce qui regarde le recensement de 1980: 2.1, 2.7 et 3.7. Comme la population visée par le TOR de 1986 était constituée en majeure partie de membres de minorités ethniques (73 pour cent de personnes d'origine hispanique, 12 pour cent de personnes d'origine asiatique et 15 pour cent qui ne sont ni d'origine asiatique ni d'origine hispanique), nous servirons des estimations d'Erickson et Kadane: $E(\theta) = 2.1, 2.7$, ou 3.7 , $\text{Var}(\theta) = 0$. Nous considérons ici que θ est fixe mais inconnu. Dans la section 6, nous réalisons une analyse de sensibilité pour illustrer l'effet de diverses valeurs de θ .

Les estimations de θ sont conformes à ce qui est indiqué dans les rapports des observateurs qui ont participé au recensement d'essai de Los Angeles (Childers et coll. 1987). Nous sommes d'avis que le biais de corrélation est plus élevé pour les régions urbaines que pour le pays en général. En conséquence, il peut s'agir ici d'estimations prudentes étant donné le caractère urbain de la région d'essai.

5.1.5 Résumé

Dans le modèle de l'erreur totale, les deux premiers moments de la distribution a posteriori du facteur de correction pour le biais de corrélation sont supposés être $E(\theta) = 2.1, 2.7$ ou 3.7 et $\text{Var}(\theta) = 0$.

L'ESD idéal peut être exprimé par la formule suivante:

$$N_{1+}N_{+1}/N_{11} = (\hat{C} - c)(\hat{N}_p - n_p)(\hat{M} - m).$$

5. COMPOSANTES DE L'ERREUR DANS L'EP

Les estimations des deux premiers moments de la distribution a posteriori du taux de sous-dénombrement découlent des estimations des deux premiers moments des composantes de l'erreur dans l'EP. Ces composantes sont le biais de corrélation, l'erreur d'appariement, l'erreur de déclaration d'adresse, la fabrication dans l'échantillon P, l'erreur dans l'estimation du nombre d'enregistrements erronés, la conciliation des estimations des taux de surdénombrement et de sous-dénombrement bruts, les données manquantes et l'erreur d'échantillonnage. Dans cette section, nous allons décrire la source de chaque composante et définir un modèle pour chacune. Aux fins de la modélisation, nous considérons les composantes d'erreur comme des indicateurs observables de la qualité des données. Nous estimons enfin les deux premiers moments des distributions des erreurs en vue de les utiliser dans le modèle de l'erreur totale défini à la section 6.

5.1 Biais de corrélation

5.1.1 Source d'erreur

Dans l'estimation de système dual, la préoccupation majeure est que l'échantillon P produise une estimation précise de la proportion de la population dénombrée lors du recensement. Le non-respect de l'une ou l'autre des hypothèses d'indépendance qui sous-tendent l'estimation de système dual peut avoir pour effet d'introduire un biais dans l'estimation de la proportion de la population dénombrée lors du recensement et, par voie de conséquence, dans l'estimation de la population.

Trois hypothèses d'indépendance sont posées pour l'estimateur de système dual: **Causalité.** Le fait qu'une personne est incluse dans la population recensée est indépendant du fait qu'elle est incluse dans l'échantillon de l'EP. Autrement dit, le rapport des produits croisés satisfait

$$\theta_i = p_{111}p_{22}/p_{112}p_{21} = 1, \text{ pour } i = 1, \dots, N.$$

Homogénéité. Les probabilités de saisie satisfont $p_{i1+} = p_{1+}$ ou $p_{i+1} = p_{+1}$ pour $i = 1, \dots, N$, dans chacune des strates formées a posteriori.

Autonomie. Le recensement et l'EP sont le résultat de N essais mutuellement indépendants. L'hypothèse de l'homogénéité est conforme au modèle composé M^h décrit dans Wolter (1986a). Tout l'appareil mathématique exposé dans Wolter (1986a) pour le modèle M_i de Peterson s'applique aussi au modèle M^h lorsqu'il y a suffisamment de données pour former des strates a posteriori où s'applique le modèle M_i .

Pour réduire le degré d'hétérogénéité de la population, le U.S. Bureau of the Census exécute une stratification a posteriori des données selon des variables démographiques et géographiques; cette méthode avait été proposée à l'origine par Sekar et Deming (1949). Elle consiste à calculer une estimation de la population dans chaque strate formée a posteriori puis à additionner toutes les estimations pour obtenir la population totale estimée. À moins que l'hypothèse de l'homogénéité ne se vérifie pas du tout, l'estimation se situe normalement entre le chiffre du recensement et le chiffre de la population réelle.

Wolter (1986b) et Cowan et Malec (1986) ont démontré que le non-respect de l'hypothèse de l'autonomie avait un effet négligeable sur le biais de l'ESD mais faisait augmenter sa variance.

Tableau 4

Code de dénombrement pour les répondants de l'échantillon P		
Echantillon P		
Code de dénombrement		
Recensé		
Enregistrement fabriqué	A ₁₁	A ₁₂
Enregistrement non fabriqué	A ₂₁	A ₂₂
Adresse au jour du recensement exacte	A ₃₁	A ₃₂
Adresse au jour du recensement inexacte		

Tableau 5

Code d'appariement pour les répondants de l'échantillon P		
Echantillon P		
Code d'appariement		
Apparié		
Enregistrement fabriqué	B ₁₁	B ₁₂
Enregistrement non fabriqué	B ₂₁	B ₂₂
Adresse au jour du recensement exacte	B ₃₁	B ₃₂
Adresse au jour du recensement inexacte		

Comme le groupe de personnes réputées incluses dans l'échantillon P selon le tableau 3 correspond au groupe de répondants qui sont réputés non fictifs d'après le tableau 4, $D_{21} = A_{21}$ et $D_{31} = A_{31}$. De plus, $A_{11} = 0$ puisqu'un enregistrement fabriqué au cours de l'EP ne pouvait exister lors du recensement. Par conséquent,

$$M = D_{11} + D_{21} + D_{31} = D_{11} + A_{21} + A_{31}.$$

Comme un enregistrement fabriqué durant l'EP n'a pas d'équivalent dans la liste du recensement, nous supposons que $B_{11} = 0$. Ainsi, $\hat{M} = B_{11} + B_{21} + B_{31} = B_{21} + B_{31}$.

L'erreur non due à l'échantillonnage dans l'estimation de N_{11} , (désignée par m , peut donc être définie comme suit:

$$\begin{aligned} m &= \hat{M} - M \\ &= (B_{11} + B_{21} + B_{31}) - (D_{11} + D_{21} + D_{31}) \\ &= -D_{11} + (B_{21} - A_{21}) + (B_{31} - A_{31}). \end{aligned}$$

L'erreur m a trois composantes: $(B_{21} - A_{21})$, qui est l'erreur introduite au cours de l'appariement (section 5.2), $(B_{31} - A_{31})$, qui est l'erreur introduite par les répondants lorsque ceux-ci indiquent mal l'adresse où ils demeureraient le jour du recensement (section 5.3); et $-D_{11}$. D_{11} a elle-même deux composantes: m_i , l'erreur attribuable à l'absence de codes d'appariement, et m_f , l'erreur due à la fabrication. La première est étudiée dans la section 5.7 tandis que la seconde fait l'objet de la section 5.4.

recensement (Wolter 1986a). Les erreurs non dues à l'échantillonnage peuvent influencer sur la précision des estimateurs de N_{+1} , de N_{1+} , et de N_{11} . Nous décrivons ci-dessous l'erreur non due à l'échantillonnage.

Dans l'estimation de N_{+1} l'erreur est définie par l'expression $\hat{C} - N_{+1} = (\hat{C} - C) + (C - N_{+1})$. Le premier terme $(\hat{C} - C)$ est l'erreur non due à l'échantillonnage nette, qui entre dans la composition du biais et de la variance, et le second terme $(C - N_{+1})$ est l'erreur d'échantillonnage, qui entre uniquement dans la composition de la variance. Définissons l'erreur non due à l'échantillonnage nette comme $c = \hat{C} - C$.

L'erreur nette c se produit durant le traitement de l'échantillon D , lorsqu'on indique que des répondants ont été enregistrés correctement ou incorrectement, selon le cas, lors du dénombrement initial alors que c est le contraire. Par conséquent, c a trois composantes : c_p , qui survient durant la collecte et le traitement des données, c_b , qui est le résultat d'un plan de sondage pour l'EP qui ne réussit pas à concilier les estimations des taux de surdénombrement et de sous-dénombrement bruts, et c_i attribuable aux données manquantes, $c = c_e + c_b + c_i$. Ces trois composantes sont étudiées dans les sections 5.5, 5.6 et 5.7 respectivement.

Dans l'estimation de N_{1+} l'erreur est définie par l'expression $\hat{N}_p - N_{1+} = (\hat{N}_p - N_p) + (N_p - N_{1+})$. Le premier terme $(\hat{N}_p - N_p)$ est l'erreur non due à l'échantillonnage, qui entre dans la composition du biais et de la variance, et le second terme $(N_p - N_{1+})$ est l'erreur d'échantillonnage, qui entre uniquement dans la composition de la variance. L'erreur non due à l'échantillonnage nette est définie $n_p = \hat{N}_p - N_p$.

L'erreur nette n_p se produit à l'étape de l'interview réservée aux personnes de l'échantillon P , plus précisément lorsque des personnes comprises dans l'échantillon P ne sont pas interviewées. Dans un tel cas, soit que des enregistrements ont été fabriqués ou qu'il y a des données manquantes. Par conséquent, n_p a deux composantes : n_{pf} , l'erreur due à la fabrication, et n_{pi} , l'erreur due aux données manquantes, $n_p = n_{pf} + n_{pi}$. La première de ces composantes est étudiée dans la section 5.3, tandis que la seconde est l'objet de la section 5.7. En ce qui concerne l'estimation de N_{11} l'erreur est définie par l'expression $\hat{M} - N_{11} = (\hat{M} - M) + (M - N_{11})$. Le premier terme $(\hat{M} - M)$ est l'erreur non due à l'échantillonnage nette, qui entre dans la composition du biais et de la variance, et le second terme $(M - N_{11})$ est l'erreur d'échantillonnage, qui entre uniquement dans la composition de la variance.

Afin de décrire plus facilement l'erreur non due à l'échantillonnage dans l'estimation de N_{11} , considérons les tableaux ci-dessous, qui portent sur les personnes comprises dans la population cible de l'échantillon P et les répondants du même échantillon. Le tableau 3 indique le nombre pondéré de personnes de la population cible de l'échantillon P dans chaque classe tandis que le tableau 4 indique le nombre pondéré de répondants de l'échantillon P dans chaque classe. Quant au tableau 5, on y trouve les estimations de l'effectif de chaque classe établies à l'aide des résultats de l'interview de l'échantillon P et de l'appariement.

Tableau 3

Population cible de l'échantillon P

Echantillon P		Code de dénombrement	
		Recensé	Non recensé
Exclus	D_{11}		D_{12}
Inclus			
Adresse au jour du recensement exacte	D_{21}		D_{22}
Adresse au jour du recensement inexacte	D_{31}		D_{32}

estimateurs de N_{+1} et N_{11} .) Par conséquent, l'estimateur a la forme suivante $N_{++} = N_p \bar{C}/M$. Le ratio \bar{C}/M renferme un facteur de correction pour les enregistrements erronés et les enregistrements qui ne contiennent pas assez d'information pour être apparés II_1 et II_2 , de manière que les cas n'ayant aucune chance d'être inclus dans le dénominateur ne figurent pas non plus dans le numérateur.

L'ESD sert à estimer le sous-dénombrement net en pourcentage, ou *taux de sous-dénombrement net*, dans le recensement,

$$S = 100 (\text{REC} - N_{++}) / N_{++}.$$

Pour la région d'essai en général (notamment le Central Los Angeles County), $\text{REC} = 355,352$, $N_p = 336,707$, $\bar{C} = 343,567$, $M = 298,204$, et $N_{++} = 388,040$. Suivant ces chiffres, le taux de sous-dénombrement net estimé est 8.42.

3. MÉTHODE D'ÉVALUATION DE L'ERREUR TOTALE

L'ESD est exposé à diverses sources d'erreur, comme des adresses inexacts dans l'échantillon P , des données manquantes (non-réponse partielle ou totale), des erreurs de réponse, des erreurs commises par l'interviewer, le biais de corrélation, l'erreur d'échantillonnage, etc. Nous chercherons ici à évaluer les effets de ces diverses sources d'erreur sur l'ESD.

La première étape consiste à exprimer l'ESD comme une fonction des composantes. Celles-ci ont été construites de telle manière que les diverses sources d'erreur agissent d'une façon indépendante ou parfaitement corrélatrice sur la plupart d'entre elles. En isolant les effets des diverses erreurs, nous pouvons plus facilement définir les principales sources d'erreur.

En second lieu, nous estimons les deux premiers moments de chaque composante d'erreur prise individuellement. Pour cela, nous nous fondons sur les résultats de diverses évaluations du TOR et de programmes de contrôle de la qualité. La manière dont les composantes ont été constituées implique que le coefficient de corrélation entre ces composantes est normalement égal à 0 ou à 1.

Nous avons eu recours à des méthodes de simulation par ordinateur pour étudier la propagation des erreurs. Nous avons supposé une distribution multidimensionnelle des composantes d'erreur, par exemple F . La définition de F était compatible avec les deux premiers moments estimés dans la section 5. Nous avons simulé la présence de composantes d'erreur par des tirages pseudo-aléatoires à même F , puis nous avons calculé l'ESD; nous avons répété cette opération 10,000 fois et la distribution empirique de l'ESD ainsi obtenue a servi d'estimation de la distribution réelle. Les deux premiers moments de cette distribution fournissent des estimations numériques de l'erreur totale de l'ESD.

Nous avons réalisé une analyse de sensibilité pour vérifier l'importance d'utiliser une forme de distribution plutôt qu'une autre pour F . Les résultats de cette analyse donnent à penser que la forme de distribution exacte (au-delà des deux premiers moments) est relativement sans importance (voir section 6).

L'analyse de l'erreur dans l'ESD s'est faite selon une approche bayésienne. Nous avons estimé les deux premiers moments des distributions des composantes d'erreur, puis nous avons déterminé la distribution a posteriori du taux de sous-dénombrement en fonction des valeurs observées de \bar{C} , N_p , M , etc.

4. COMPOSANTES DE L'ESD

L'ESD est exposé aux erreurs d'échantillonnage et aux erreurs non dues à l'échantillonnage, y compris la non-vérification des hypothèses qui sous-tendent le modèle de l'ESD. Cet estimateur comporte effectivement un biais mais celui-ci est négligeable en ce qui concerne le

de f comprend 0 et 1, on peut choisir aussi bien REC que ESD. Les critères permettant de juger de la supériorité d'une série d'estimations démographiques par rapport à une autre peuvent être fondés sur des mesures de la qualité de la distribution de la population (Hogan et Mulry 1987; Spencer 1986). Les estimations de l'erreur totale dans l'ESD jouent aussi un rôle important dans la planification statistique (par ex., combien d'argent devrait être dépensé et quelle

devrait être la taille d'un échantillon de l'EP.)

Les ESD sont soumis à plusieurs composantes de l'erreur non due à l'échantillonnage, sans compter l'erreur d'échantillonnage. Nous présentons dans cet article des modèles de l'erreur totale et des composantes d'erreur dans l'ESD. Ces modèles établissent un rapport entre des indicateurs observés de la qualité des données et les deux premiers moments des composantes d'erreur. Nous utilisons ensuite des méthodes de propagation de l'erreur pour estimer le biais et la variance de l'ESD. De cette façon, nous évaluons l'erreur totale ou l'effet combiné des erreurs. Parmi les ouvrages qui abordent les modèles d'erreur pour l'ESD, voir Selitzer et Adlakha (1974).

La méthode exposée est appliquée au recensement du Central Los Angeles County de 1986, connu aussi sous le nom de Test des opérations de redressement (TOR) de 1986 pour la région de Los Angeles (Ditford 1988). L'échantillon de l'EP réalisée à l'occasion du TOR comprenait environ 6,000 unités de logement et plus de 19,000 personnes. Une analyse de sensibilité montre comment les composantes d'erreur influent l'une sur l'autre, lesquelles s'annulent et lesquelles se complètent. Les méthodes exposées ici pour estimer l'erreur dans l'ESD du TOR peuvent aussi servir à estimer l'erreur dans les ESD du recensement de 1990.

Nous avons cherché à structurer cet article de manière à faciliter la tâche de ceux qui ne veulent pas le lire en entier. Ainsi dans la section 2, nous exposons les fondements de l'ESD utilisé dans le TOR de même que ses principales composantes. Nous décrivons ensuite la méthode utilisée pour évaluer les composantes d'erreur et les combiner dans le but d'estimer l'erreur totale dans l'ESD (Section 3). Une description précise des composantes d'erreur exige une description détaillée de l'ESD, appuyée de symboles (Section 4). Cette description est suivie d'une évaluation des composantes d'erreur (Section 5). Une synthèse de ces composantes aboutit au calcul d'estimations de l'erreur totale dans l'ESD (Section 6). Enfin, nous présentons les principales conclusions (Section 7).

2. ESTIMATEUR DE SYSTÈME DUAL

L'utilisation de l'estimateur de système dual suppose nécessairement qu'il existe deux listes de la population. La première est le résultat du dénombrement initial et la seconde est une liste implicite des personnes incluses dans la base de sondage servant à la formation de l'échantillon P de l'EP, et qui constituent ce que nous appellerons la population de l'échantillon P. La base de sondage proprement dite n'est pas une liste de personnes mais d'îlots de recensement. L'échantillon P est un des deux échantillons utilisés dans l'EP. Celle-ci comprend l'échantillon D, qui est constitué d'enregistrements du recensement, et l'échantillon P, qui est un

Probabilité d'inclusion dans une case

Recensement			
Inclus		Exclus	
Echantillon P inclus		Total	
p_{i11}	p_{i+1}	p_{i+2}	p_{i++}
p_{i21}		p_{i22}	p_{i2+}
p_{i1+}			p_{i++}

Tableau 1

L'erreur totale dans l'estimateur de système dual: Recensement du Central Los Angeles County de 1986

MARY H. MULRY et BRUCE D. SPENCER¹

RÉSUMÉ

Le U.S. Bureau of the Census utilise des estimateurs de système dual (ESD) pour évaluer l'erreur de couverture dans le recensement. Ce genre d'estimateur repose sur des données du recensement initial et d'une enquête postcensitaire. Lorsque l'on mesure la précision de l'ESD, il importe de savoir que cet estimateur est soumis à plusieurs composantes de l'erreur d'échantillonnage et de l'erreur non due à l'échantillonnage. Dans cet article, nous décrivons des modèles de l'erreur totale et des composantes d'erreur dans les estimateurs de système dual. Ces modèles établissent un rapport entre des indices observés de la qualité des données, comme le taux d'erreur d'appariement, et les deux premiers moments des composantes d'erreur. Nous analysons également la propagation de l'erreur dans l'ESD et évaluons le biais et la variance de cet estimateur. La méthode proposée est appliquée au recensement du Central Los Angeles County de 1986 dans le cadre du Test des opérations de redressement du U.S. Bureau of the Census. Cette méthode sera aussi utile pour évaluer l'erreur dans l'ESD à l'occasion du recensement de 1990 et pour d'autres applications.

MOTS CLÉS: Erreur non due à l'échantillonnage; enquête postcensitaire; évaluation de la couverture; sous-dénombrement; saisie-résaisie.

1. INTRODUCTION

L'estimateur de système dual (ESD) est utilisé dans plusieurs disciplines pour estimer la taille d'une population. Il peut s'agir de populations d'animaux comme de populations d'êtres humains. Par exemple, le U.S. Bureau of the Census utilise des ESD du nombre de naissances dans son analyse démographique pour estimer la population des États-Unis. Il prévoit utiliser le même genre d'estimateur pour mesurer l'erreur de couverture dans le recensement décennal de 1990. Dans cet article, nous allons insister sur l'application de l'ESD dans le domaine du recensement, où les deux systèmes sont le dénombrement initial et une enquête postcensitaire (EP).

L'estimateur le plus élémentaire fondé sur l'ESD du sous-dénombrement dans le recensement est \hat{SD} et est défini $\hat{SD} = ESD - REC$, où REC désigne le chiffre du recensement. Puisque $ESD = REC + \hat{SD}$, les ESD produisent aussi d'autres estimateurs de population. Une catégorie plus générale d'estimateurs fondés sur l'ESD (Spencer 1980, 1986) est $(1 - f) \times REC + f \times ESD$, qui équivaut à

$$REC + f \times UC$$

où $0 \leq f \leq 1$.

Les estimations de l'erreur totale dans l'ESD sont indispensables pour déterminer quelle valeur de f produit l'estimateur de population le plus précis. Comme l'intervalle des valeurs

¹ Mary Mulry, Undercount Research Staff, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C. 20233. Bruce Spencer, Department of Statistics, Northwestern University, Evanston, IL 60201 et NORC. Les opinions exprimées dans cet article sont celles des auteurs et ne reflètent pas nécessairement la position du U.S. Bureau of the Census.

BIBLIOGRAPHIE

- BAILLAR, B., et MARTIN, E. (1987). Report on Meetings in Los Angeles, Chicago and Denver. Note de service inédite du Bureau of the Census.
- CHOLDIN, H. (1987). Science and Scientists in the 1980 Census Law suits. Document présenté à la réunion de mai 1987 de la Population Association of America, Chicago, Ill.
- CLOGG, C.C., MASSAGLI, M.P., et ELIASON, S.R. (1986). Population undercount as an issue in social research". *Proceedings of the Second Annual Research Conference, March 23-26*. United States Bureau of the Census, Washington, D.C., 335 — 343.
- CITRO, C.F., et COHEN, M.L. (éds.) (1985). *The Bicentennial Census: New Directions for Methodology in 1990*, colloque sur la méthodologie des recensement décennaux, National Research Council, Washington, D.C., National Academy Press.
- DIFFENDAL, G. (1988). Test des opérations de redressement de 1986 dans le Central Los Angeles County. *Techniques d'enquête*, 14, 75-92.
- DILLMAN, D.A. (1978). *Mail and Telephone Surveys: The Total Design Method*, New York: John Wiley and Sons.
- EDSON, R.G. (1987). Preliminary coverage improvement results from tests for the 1990 Census. Document présenté à la réunion d'août 1987 de l'American Statistical Association, San Francisco.
- ERICKSEN, E.P. (1983). Affidavit, document présenté au District Court, Southern District of New York, Mario Cuomo et ass., Plaintiff(s), Malcolm Baldrige et ass., Defendants, 80 Civ. 4550 (JES).
- FAY, R.E., PASSEL, J.S., et ROBINSON, J.G. (1988). The coverage of population in the 1980 census. *Evaluation and Research Reports. 1980 Census of Population and Housing*, PH80-E4, Washington, D.C.
- HAINER, P., HINES, C., MARTIN, E., et SHAPIRO, G.M. (1988). Research on improving coverage in household surveys. *Proceedings of the Fourth Annual Research Conference*. United States Bureau of the Census, Washington, D.C., 513 à 539.
- HEER, D.M., et PASSEL, J.S. (1987). Comparison of two methods for estimating the number of undocumented Mexican adults in Los Angeles County. *International Migration Review*, 21(4), 1446 — 1473.
- HOGAN, H., et WOLTER, K. (1988). Mesure de l'erreur dans une enquête postcensitaire. *Techniques d'enquête*, 14, 105 — 124.
- KEYFITZ, N. (1979). Information and allocation: Two uses of the 1980 census. *The American Statistician*, 33(2), 45 — 56.
- MOORE, J.C., et McDONALD, S.K. (1987). The Census community awareness program: an evaluation of the potential and actual effectiveness of CCAP based on evidence from the 1986 Los Angeles Census Test. Rapport inédit du Bureau of the Census.
- ORTNER, R. (1987). Statement. *United States Department of Commerce News*, 30 octobre 1987.
- U.S. BUREAU OF THE CENSUS (1960). *The Post Enumeration Survey: 1950*. Rapport technique n° 4, Washington, D.C.
- U.S. BUREAU OF THE CENSUS (1987a). 1986 Test Census, Central Los Angeles County, California. *General population and Housing Statistics*, TC86-1, Washington, D.C.
- U.S. BUREAU OF THE CENSUS (1987b). Programs to improve coverage in the 1980 census. *Evaluation and Research Reports. 1980 Census of Population and Housing*, PHC80-E3, Washington, D.C.
- U.S. BUREAU OF THE CENSUS (1987c). *Statistical Abstract of the United States*: 1987, (106e édition), Washington, D.C.
- U.S. GENERAL ACCOUNTING OFFICE (1980). *Problems in Developing the 1980 Census Mail List*. Washington, D.C.: General Accounting Office.

On ne saurait appliquer sans précaution les conclusions de la présente étude sur les causes du sous-dénombrement au recensement des régions urbaines en 1990. En effet, celles-ci étant fondées sur les résultats d'un recensement d'essai, les erreurs observées peuvent résulter d'un manque d'expérience dû au caractère expérimental des procédures appliquées ou à la difficulté de convaincre les répondants (et le personnel chargé d'effectuer le dénombrement) que le projet est aussi important qu'un recensement décennal. En outre, étant donné que Los Angeles présente des caractéristiques bien différentes de celles des autres grandes régions urbaines, les problèmes que pose le dénombrement dans cette ville peuvent lui être particuliers. Par exemple, Los Angeles passe pour abriter plus d'immigrants illégaux que toute autre grande ville des États-Unis (Heer et Passel 1987).

Par ailleurs, le taux de sous-dénombrement net enregistré pour Los Angeles en 1980 est semblable à celui des autres grandes villes, comme l'a révélé le programme postcensitaire de 1980 (Fay et coll. 1988). Autrement dit, si les autres villes abritent moins d'immigrants illégaux, elles comptent probablement davantage de sous-populations difficiles à dénombrer. Il conviendrait d'entreprendre une recherche plus détaillée afin de déterminer dans quelle mesure les causes du sous-dénombrement diffèrent selon la race, l'origine ethnique et d'autres caractéristiques sociales.

Il est encourageant de constater que les causes du sous-dénombrement mises en évidence dans l'étude fondée sur les résultats de l'enquête postcensitaire corroborent dans une bonne mesure les conclusions des rapports plus qualitatifs produits par des ethnographes et des groupes intéressés à la question. En outre, les estimations du sous-dénombrement établies dans le cadre de l'enquête postcensitaire à partir des résultats du recensement d'essai effectué à Los Angeles sont considérées comme étant de grande qualité (Hogan et Wolter, 1988). Pour ces diverses raisons, il est recommandé d'étendre aux autres régions urbaines (et non urbaines) la méthode fondée sur l'enquête postcensitaire qui est décrite dans le présent document.

En ce qui a trait au contexte social, il serait utile, pour pouvoir déterminer plus adéquatement jusqu'à quel point la couverture du recensement peut être améliorée au moyen des programmes de sensibilisation du public et d'action communautaire mis sur pied par le Bureau of the Census, de procéder à un complètement de recherche en vue d'établir de quelle façon les recenseurs évaluent les coûts et les avantages que représente pour eux leur participation au recensement. Il conviendrait également d'avoir de meilleurs indicateurs des raisons qui poussent les répondants à oublier délibérément certains membres du ménage. Un examen des programmes d'aide permettrait de confirmer l'incidence possible de la participation au recensement sur l'admissibilité aux programmes de bien-être car, l'admissibilité aux programmes d'aide n'est pas nécessairement mise en danger par la déclaration de la composition véritable du ménage.

Il faut également prévoir d'améliorer l'évaluation des causes du sous-dénombrement résultant des opérations mêmes de dénombrement. En combinant les données tirées des programmes de contrôle de la qualité des données du recensement et les résultats de l'appariement effectué dans le cadre de l'enquête postcensitaire, il serait possible d'identifier les sources d'erreurs avec plus de précision.

REMERCIEMENTS

Nous tenons à remercier les personnes suivantes pour leur contribution à cette étude: Irwin Anolik, Miriam Balutis, Gregg Diffendal, Chris Dyke, Sue Finnegan, Howard Hogan, Jan Jaworski, Pete Long et Lynn Weidman. Betsy Martin, Jim O'Brien et deux relecteurs anonymes ont fourni des commentaires utiles sur les versions intermédiaires du présent document.

problèmes pourraient améliorer la situation dans une certaine mesure. On pourrait également avoir recours à des procédures spéciales de rappel dans le cas des ménages de petite taille, des ménages plus mobiles et de ceux dont les membres sont rarement à la maison.

Il est certain que ces améliorations apportées à un recensement qui est déjà considéré comme un grand succès seront coûteuses. Keyfitz (1979) et d'autres considèrent que l'augmentation des coûts qui résulterait du fait d'ajouter des personnes est d'autant plus forte que le taux de couverture obtenu approche de 100%. Les modifications apportées aux programmes en vue de réduire les erreurs observées au recensement d'essai de 1986 s'ajouteraient aux 2,6 milliards de dollars, montant prévu des coûts du recensement de 1990, étant donné que la méthode qui sera utilisée dans les régions urbaines sera analogue à celle qui a été appliquée lors du recensement d'essai effectué à Los Angeles.

L'oubli de personnes à l'intérieur du ménage est un problème plus difficile à résoudre que celui de l'oubli de ménages. Le Bureau of the Census doit redoubler ses efforts en vue de mieux comprendre les situations complexes de certains particuliers à l'intérieur des ménages et les facteurs cognitifs et (ou) culturels qui influent sur la perception que les gens ont des liens à l'intérieur du ménage. Les résultats de la présente étude laissent entrevoir que l'on pourrait réduire les erreurs d'interprétation des définitions en portant une attention particulière aux répondants pour qui l'anglais n'est pas la langue maternelle et aux ménages composés de personnes qui ont des liens éloignés.

Cependant, compte tenu de l'importance de la recherche effectuée en vue d'améliorer la conception du questionnaire de recensement et les règles complexes relatives au lieu de résidence et au dénombrement que le Bureau of the Census doit maintenir en raison même de son statut et des antécédents, l'application de mesures complémentaires visant à réduire les erreurs d'interprétation des définitions nécessiterait des efforts extraordinaires. Ces erreurs reposent largement sur les différences culturelles et l'insuffisance de scolarité des sous-populations difficiles à dénombrer.

L'oubli de personnes à l'intérieur du ménage est également apparu comme étant fortement relié à la présence d'immigrants récents et de bénéficiaires du bien-être ainsi qu'à la densité d'occupation du logement. Le fait que l'étude effectuée à partir des résultats de l'enquête postcensitaire mette en évidence l'incidence de ces variables indique qu'à l'occasion de cette dernière, on a réussi à dénombrer bon nombre de personnes qui avaient été oubliées lors du recensement. L'incidence des variables liées à la non-déclaration délibérée de certaines personnes dans le ménage peut, dans une certaine mesure, être attribuée également à des facteurs non contrôlés autres que la non-déclaration délibérée mais la persistance de ces liens même lorsque l'on inclut la variable "composition du ménage" dans le modèle log-linéaire (non illustré ici) indique que l'enquête postcensitaire a réellement permis de dénombrer un certain nombre de personnes qui avaient été oubliées au moment du recensement. En d'autres mots, il semble y avoir un certain continuum entre les ménages très réfractaires au dénombrement et ceux qui le sont moins et, dans le cas de ces derniers, l'application de méthodes plus directement incitatives, comme celles qui ont été utilisées pour l'enquête postcensitaire, peut donner de bons résultats. Les conditions sociales qui caractérisent les cas les plus irrévocables de non-déclaration délibérée de certaines personnes dans le ménage constituent un problème majeur pour l'U.S. Bureau of the Census. Les programmes de sensibilisation du public visant à convaincre la population de l'importance du recensement et du fait que le caractère confidentiel des données sera respecté se sont révélés sans effet sur la sous-population particulièrement difficile à dénombrer qui était visée par le recensement d'essai effectué à Los Angeles, comme cela est décrit par Moore et McDonald (1987), encore que ces programmes pourraient avoir de meilleurs résultats dans le contexte du véritable recensement décennal. La très faible incidence de la conviction du répondant vis-à-vis du respect de la confidentialité des données, considérée en elle-même ou comme moyen de concilier les conditions propres du ménage et les craintes du répondant qui préfère ne pas déclarer certains membres du ménage, laisse entrevoir que le lien entre les réactions individuelles et la décision de participer au recensement est loin d'être simple.

Les personnes oubliées au moment du recensement dans les ménages de la catégorie "non-appartient partiel" sont un peu plus susceptibles que celles qui ont été oubliées dans les ménages de la catégorie "appartient partiel", d'être de sexe masculin et de n'être pas scolarisées et un peu moins susceptibles d'avoir la citoyenneté américaine ou d'être proches parentes de la personne responsable du ménage (tableau 7). Les personnes oubliées au moment du recensement dans les ménages de la catégorie "non-appartient partiel" sont également légèrement plus susceptibles de n'être pas citoyens américains et d'avoir un niveau de scolarité moindre que celles qui ont été oubliées dans des ménages de la catégorie "appartient partiel", mais, dans leur cas, on ne relève aucune incidence du sexe ni du lien avec la personne responsable du ménage. Ainsi, dans l'ensemble, les personnes oubliées dans les ménages de la catégorie "non-appartient partiel" diffèrent de celles qui ont été oubliées dans des ménages de la catégorie "appartient partiel" sous d'avantage de rapports que les personnes oubliées dans des ménages de la catégorie "non-appartient partiel".

En plus de baisser un plus grand nombre de caractéristiques du recensement, l'oubli de personnes dans des ménages de la catégorie "non-appartient partiel" a entraîné l'oubli d'un nombre beaucoup plus important de personnes que ce n'est le cas dans les ménages de la catégorie "non-appartient partiel". Sur l'ensemble des cas de non-appartient partiel de l'échantillon de l'enquête postcensitaire, les deux tiers (67%) correspondent à des ménages de la catégorie "non-appartient partiel" et un tiers seulement à des ménages de la catégorie "appartient partiel". En tout, 82% des personnes oubliées ont été repérées lors de l'enquête postcensitaire dans des logements dûment dénombrés au recensement et 18% seulement dans des logements oubliés au recensement.

5. DISCUSSION

Les résultats exposés précédemment corroborent les observations tirées d'autres études qualitatives selon lesquelles, à l'heure actuelle, l'oubli de personnes dans des ménages de la catégorie "non-appartient partiel" est la cause majeure du sous-dénombrement dans les régions urbaines particulièrement difficiles à dénombrer des États-Unis. Par rapport aux personnes oubliées dans les ménages, complètement oubliées les personnes oubliées dans des ménages de la catégorie "non-appartient partiel" lors du recensement d'essai effectué à Los Angeles sont deux fois plus nombreuses, elles sont le fait de causes que l'on peut plus difficilement déterminer et elles contribuent à biaiser plus largement les données sur les caractéristiques des personnes.

Les principaux problèmes reliés à l'oubli de personnes dans des ménages de la catégorie "non-appartient partiel" sont l'oubli de certains types de logement au moment de l'établissement des listes d'adresses pour le recensement et la classification erronée de logements occupés comme logements inoccupés. Les logements qui risquent le plus d'être classifiés par erreur comme logements inoccupés sont ceux qui sont occupés par des ménages de petite taille très mobiles et ceux dont tous les membres adultes travaillent à plein temps. Les résultats des programmes d'amélioration de la couverture du recensement mis sur pied par le Bureau of the Census laissent entrevoir que le nombre de logements oubliés pourrait être réduit. C'est grâce à ces programmes que l'on a pu rajouter 10% environ de logements sur les listes d'adresses lors du recensement d'essai effectué à Los Angeles. À l'occasion du recensement d'essai, le Bureau of the Census a appliqué des procédures de prédénombrement afin de repérer ces logements dans les grands immeubles à logements multiples. La figure 2 fait état de la mesure importante dans laquelle on a pu réduire les effets de cette source d'erreur au recensement d'essai: aucun des logements oubliés n'était situé dans ce type d'immeubles.

Il est plus difficile d'éviter la classification erronée de logements occupés comme logements inoccupés. Le fait d'accorder davantage de temps aux intervieweurs pour le suivi de chaque cas de non-réponse de même qu'une formation plus détaillée en ce qui a trait à certains

Estimations des paramètres pour les interactions entre les indicateurs de la non-déclaration
délibérée de certaines personnes dans le ménage et les indicateurs de l'oubli de personnes
dans des ménages de la catégorie "non-apparemment parti(e)" dans le modèle
final relatif à la non-déclaration délibérée de certaines personnes

Fréquence marginale: ménages de la catégorie "non-appariement partiel" et . . .	Estimation du paramètre	Erreur- type	Estimation normalisée du paramètre
Présence d'immigrants récents:	.19	.06	3.2
Immigrants présents			
Présence de bénéficiaires du bien-être:	.17	.05	3.4
Bénéficiaires de prestations présents			
Densité d'occupation:			
Moins de .5 personne par pièce	-.49	.13	-3.8
De .5 à 1 personne par pièce	-.01	.08	-.1
De 1 à 1.5 personne par pièce	.08	.08	1.0

Répartition en pourcentage des caractéristiques des personnes selon le statut d'appariement lors de l'enquête postcensitaire et le type de ménage

Statut d'appariement lors de l'enquête postcensitaire		Caractéristiques	
Appariement		Appariement intégral du ménage HHS	du ménage HHS
Non-appariement		Non-appariement partiel du ménage HHS	Non-appariement partiel du ménage HHS

Sexe	Masculin	Féminin	<i>n</i> (non pondéré)	Niveau de scolarité	Aucune scolarité	Scolarité inférieure au secondaire	Études secondaires partielles	Diplôme d'études secondaires	<i>n</i> (non pondéré)	Lien avec le responsable du ménage	Personne appartenée faisant partie de la famille nucléaire	Personne appartenée ne faisant pas partie de la famille nucléaire	Personne non appartenée	<i>n</i> (non pondéré)	Citoyenneté	Citoyen américain de naissance	Citoyen américain par naturalisation	Citoyen non américain	<i>n</i> (non pondéré)
46.2%	53.8	49.4	2564	10.9	17.0	30.7	20.5	38.6	1197	86.1	11.3	12.6	2.6	1659	66.2	53.5	9.5	24.6	1223
50.6%	49.4	45.9	1324	10.9	17.0	34.4	20.6	34.1	1560	83.2	85.9	63.6	25.4	2560	52.6	50.4	6.4	37.0	1567
54.2%	51.8	58.2	582	14.3	37.5	30.7	19.5	28.8	599	85.9	7.9	25.4	11.0	1359	50.4	50.4	6.4	41.0	612
48.2%	48.2%	51.8	582	14.3	37.5	30.7	19.5	28.8	315	85.9	7.9	25.4	11.0	590	50.4	50.4	6.4	43.2	316

Il convient de souligner que les liens entre l'oubli de personnes dans les ménages de la catégorie "non-appariement partiel" et la taille du ménage ne tiennent plus lorsque l'on fait entrer en jeu la densité d'occupation du logement (voir le tableau 5). Cela signifie que la taille du ménage n'a pas de véritable incidence par elle-même mais qu'elle en a une seulement à cause de sa relation avec la densité d'occupation du logement. L'incidence de la densité d'occupation du logement est elle-même fortement liée à la présence d'immigrants récents dans le ménage.

4.2 Caractéristiques des personnes

Dans la dernière partie de l'analyse, qui porte sur le lien entre les caractéristiques des personnes et le sous-dénombrement, on compare quatre types de personnes : les personnes qui, lors du recensement, ont été dénombrées dans des ménages classés dans les catégories "appariement intégral", "non-appariement partiel", et "non-appariement intégral". La figure 6 illustre les écarts entre les pourcentages correspondant à dix groupes d'âge, pour les personnes appartenant à des ménages de la catégorie "appariement intégral" et à chacun des trois autres groupes. Elle révèle un pourcentage particulièrement élevé de personnes oubliées dans le groupe des personnes âgées de 20 à 29 ans relativement aux ménages de la catégorie "appariement intégral" pour les ménages des catégories "non-appariement partiel" et "non-appariement intégral" ainsi que pour les personnes du même groupe d'âge dénombrées dans des ménages de la catégorie "non-appariement partiel".

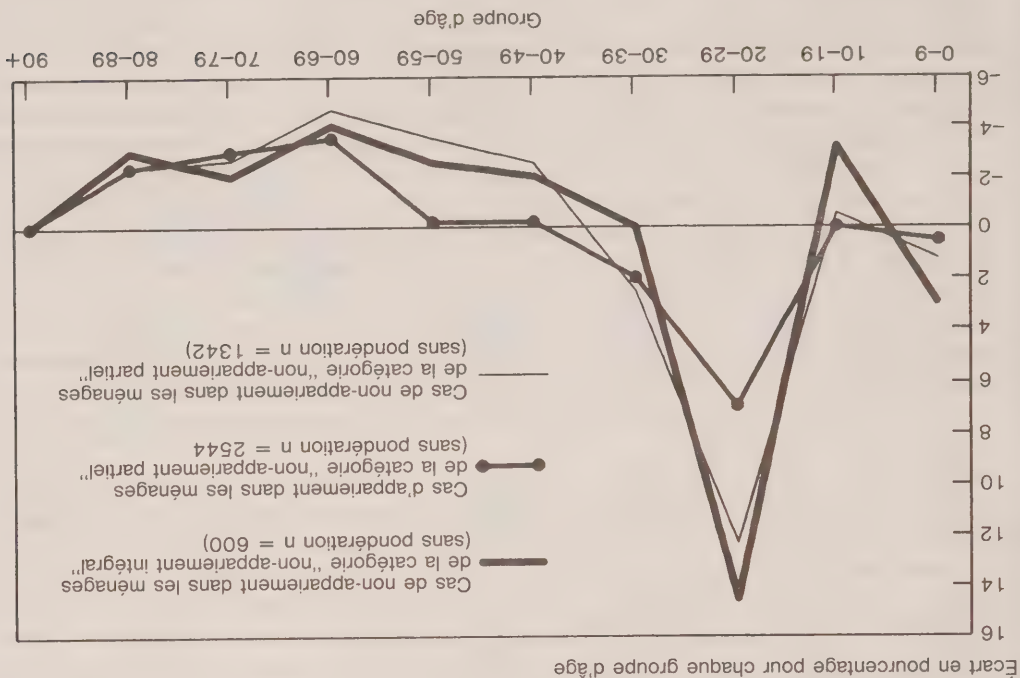


Figure 6. Écart en pourcentage, pondéré, pour chaque groupe d'âge relatif aux membres des ménages de la catégorie "non-appariement partiel".

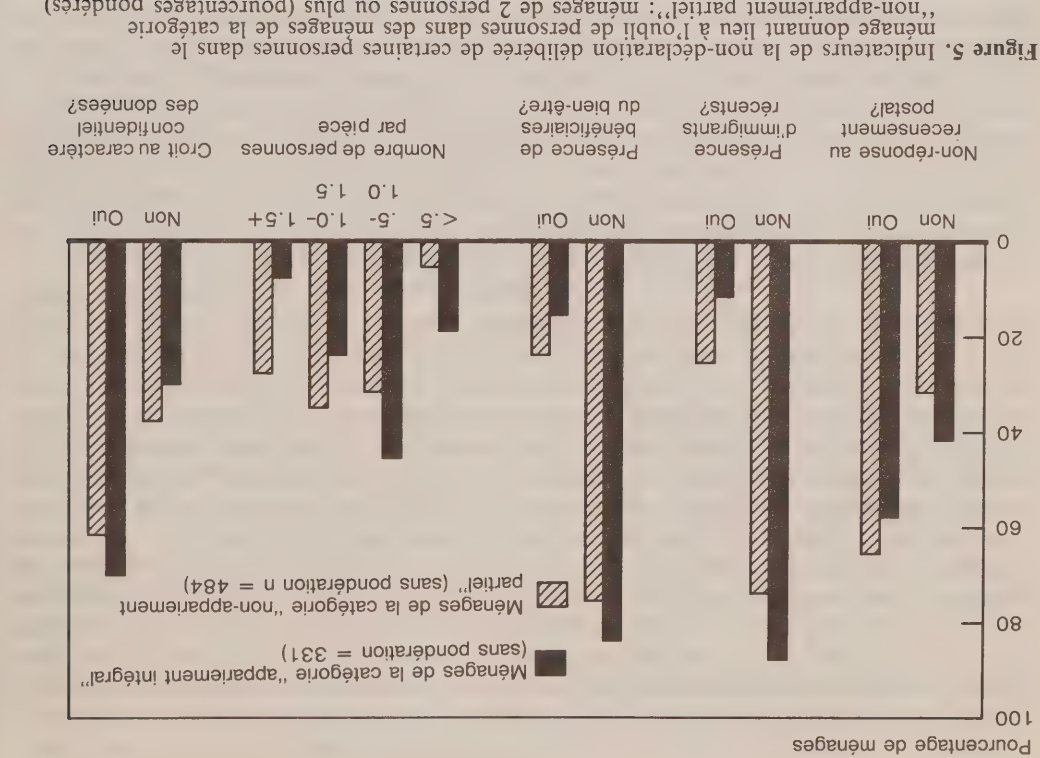


Tableau 5

Statistiques Khi carré utilisées pour détecter des interactions entre deux critères de classification dans le modèle final relatif à la non-déclaration délibérée de certaines personnes dans le ménage^a

Variables				
Densité d'occupation	Présence de bénéficiaires du bien-être	Présence d'immigrants récents	Taille du ménage	Oubli de personnes
16.7**	10.1**	11.3**	2.9	
221.7**	7.5*	.2	-	
30.0**	1.6	-	-	
5.4	-	-	-	Présence de bénéficiaires du bien-être

Interaction avec . . .

**: p < .01
*: p < .05
^a Khi carré de Wald = 103.8, df = 150, p = .9985.

Indicateurs de non-déclaration délibérée

Les facteurs que l'on suppose être à l'origine de la non-déclaration délibérée de certains membres du ménage par les répondants au recensement incluent notamment la crainte que des personnes entrées illégalement au pays soient expulsées, que, par suite de la déclaration de la présence d'adultes de sexe masculin, les prestations de bien-être soient supprimées et que le fait de révéler que le nombre d'occupants du logement dépasse le nombre autorisé entraîne des complications avec le propriétaire. Les indicateurs correspondants sont, respectivement, la présence d'immigrants récents parmi les membres du ménage, c'est-à-dire de personnes entrées au pays en 1980 ou après cette date, la présence dans le ménage de bénéficiaires du bien-être touchant des prestations pour le mois durant lequel le recensement est effectué et le nombre moyen de personnes par pièce. La non-réponse au recensement postal est également considérée comme un indicateur général d'une propension chez les recensés à considérer qu'ils n'ont rien de bon à attendre de leur participation au recensement. Enfin, on a tenu compte de la mesure dans laquelle les recensés sont prêts à croire que la confidentialité des données sera respectée afin de déterminer si cet aspect contribue à réduire les craintes à l'origine de la non-déclaration délibérée de certaines personnes dans le ménage.

La figure 5 révèle que, au niveau de l'analyse à deux variables, tous ces indicateurs sont liés à l'oubli de personnes dans des ménages de la catégorie "non-appariement partiel". Ainsi, 26% des ménages classés dans cette catégorie (trait léger) comptaient des immigrants récents, le pourcentage correspondant pour la catégorie "appariement intégral" (trait foncé) étant de 12% seulement. De même, 24% des ménages de la première catégorie ont déclaré avoir parmi leurs membres des bénéficiaires du bien-être par opposition à 15% seulement pour les ménages de la catégorie "appariement intégral". Les ménages de la catégorie "non-appariement partiel" sont davantage susceptibles que ceux de la catégorie "appariement intégral" de se caractériser par une forte densité d'occupation du logement; 63% comptaient plus d'une personne par pièce, par opposition à 34% seulement dans le cas des ménages classés dans la seconde catégorie. Les premiers sont également moins susceptibles que les seconds d'avoir retourné leur questionnaire par la poste et de croire que le caractère confidentiel des données sera respecté. Cette fois encore, on a ajusté les modèles log-linéaires, en utilisant l'oubli de personnes dans les ménages de la catégorie "non-appariement partiel" comme variable dépendante et les indicateurs de la non-déclaration délibérée de certaines personnes dans le ménage comme variables indépendantes. Les cas d'interaction de deux critères de classification avec la taille du ménage ont été inclus à titre de contrôle étant donné que, toutes choses étant égales d'ailleurs, les ménages de plus grande taille sont davantage susceptibles que ceux qui sont de plus petite taille de se caractériser par une forte densité d'occupation du logement et de compter des immigrants récents.

Cette fois, par contre, deux variables ont été rejetées aux essais préliminaires: la non-réponse au recensement postal et la conviction que le caractère confidentiel des données sera respecté. Avant de rejeter définitivement cette dernière variable, on a effectué des essais complémentaires pour déterminer si l'interaction sur les cas d'oubli de personnes dans les ménages de la catégorie "non-appariement partiel" de variables telles que la présence dans le ménage d'immigrants récents ou de bénéficiaires du bien-être, ou une forte densité d'occupation du logement dépend de la conviction que le caractère confidentiel des données sera ou ne sera pas respecté. Toutefois, il est apparu que ce dernier aspect n'influit pas sur les interactions observées.

Le tableau 5 montre qu'il y a interaction des trois variables restantes, en ce qui a trait à la non-déclaration délibérée de certaines personnes dans le ménage, c'est-à-dire la présence d'immigrants récents ou de bénéficiaires du bien-être et une forte densité d'occupation du logement, sur l'oubli de personnes dans les ménages de la catégorie "non-appariement partiel" dans un modèle incluant toutes les interactions avec la taille du ménage, d'une part, et toutes les interactions entre les variables indépendantes prises deux à deux, d'autre part. Les estimations des paramètres normalisés (voir le tableau 6) témoignent que ces trois indicateurs ont sensiblement les mêmes effets.

L'analyse à plusieurs variables révèle des interactions significatives avec l'oubli de personnes dans les ménages de la catégorie "non-appariement partiel" pour tous les indicateurs d'erreurs d'interprétation des définitions, à l'exception du niveau de scolarité du répondant. Le tableau 3 présente les statistiques Khi carré (Wald) associées au modèle final relatif à l'erreur d'interprétation des définitions qui ne fait pas état du niveau de scolarité du répondant. On relève également des interactions significatives entre la taille du ménage et sa composition ainsi que le fait que la langue parlée à la maison soit autre que l'anglais. Les estimations des paramètres présentées dans le tableau 4 révèlent que ces interactions vont bien dans le sens prévu. Les estimations pour les paramètres normalisés, obtenues en divisant les estimations pour le paramètre par l'erreur-type correspondante, indiquent que l'incidence de la taille et de la composition du ménage sont d'importance analogue et que, dans les deux cas, elle est supérieure à celle du suivi après vérification et de la langue parlée à la maison autre que l'anglais.

Tableau 3

Statistiques Khi carré utilisées pour détecter des interactions entre deux critères de classification dans le modèle final relatif aux erreurs d'interprétation des définitions^a

Variable	Interaction avec . . .			
	Langue parlée à la maison	Suivi après vérification	Composition du ménage	Taille du ménage
Oubli de personnes	38.1**	42.3**	112.0**	6.3*
Taille du ménage	-	9	50.0**	1.3
Composition du ménage	-	1.6	-	-
Envoyé pour un suivi après vérification	-	-	-	1.0

** : p < .01
 * : p < .05
^a Khi carré de Wald = 42.2, df = 45, p = .5922.

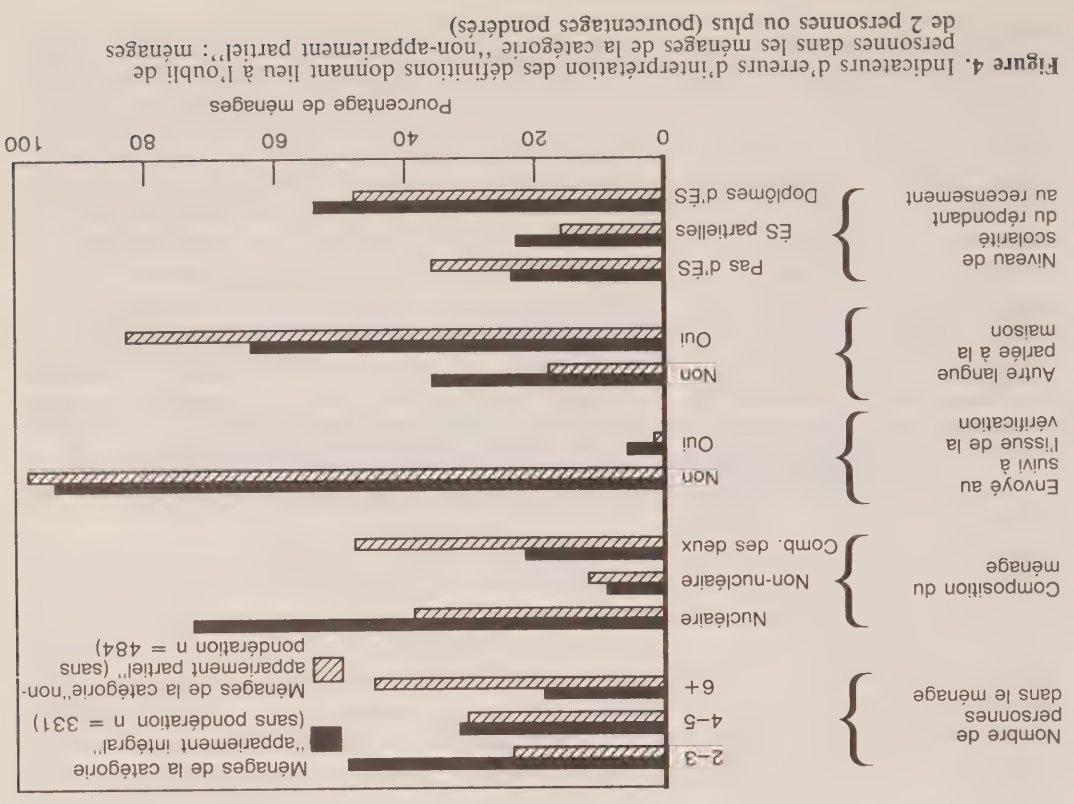
Tableau 4

Estimations des paramètres pour les interactions entre les indicateurs d'erreurs d'interprétation des définitions et les indicateurs de l'oubli de personnes dans des ménages de la catégorie "non-appariement partiel" dans le modèle final

Fréquence marginale:	Estimation du paramètre	Erreur-type	Estimation normalisée du paramètre
Taille du ménage:			
2 à 3 personnes	-.34	.06	-5.7
4 à 5 personnes	-.02	.05	-.4
Composition du ménage:			
Famille nucléaire	-.36	.06	-6.0
Autre	.22	.09	2.4
Envoyé pour un suivi après vérification	.25	.10	2.5
Non envoyé pour le suivi	.10	.05	2.0
Autre langue parlée à la maison?			
Oui			

que ceux de la catégorie "appariement intégral" (64%). Enfin, les répondants des ménages de la catégorie "non-appariement partiel" ont un niveau de scolarité moindre que ceux des ménages de la catégorie "appariement intégral": dans le premier cas, 36% des répondants au recensement n'ont pas fréquenté l'école secondaire, le pourcentage correspondant étant de 24% dans le second cas.

On a alors ajusté les modèles log-linéaires afin de déterminer si ces différences persistent lorsque l'on prend en considération simultanément plusieurs variables. Dans ces modèles, la variable dépendante est l'oubli de personnes dans des ménages de la catégorie "non-appariement partiel", les ménages de la catégorie "appariement intégral" étant codés "0" et ceux de la catégorie "non-appariement partiel" étant codés "1". On effectue une série de modèles à classification hiérarchique pour vérifier les interactions entre l'oubli de personnes dans les ménages de la catégorie "non-appariement partiel" et chacune des variables indépendantes de la figure 4. Toutes les interactions entre deux critères de classification, ont été incluses dans chacun des modèles pour des fins de contrôle.



moyenne de taille beaucoup plus petite que les ménages dont une partie ou la totalité des membres a été dénombrée (trait foncé). Alors que 53% de l'ensemble des ménages vivant dans des logements dument dénombrés qui n'ont pu être apparés comptaient un ou deux membres, 35% seulement des ménages dument dénombrés étaient de cette taille.

Les indicateurs de la propension à déménager des ménages sont le statut à l'égard de la propriété et la mobilité du ménage au cours des quatre mois écoulés entre le recensement et l'enquête poscensitaire. Les ménages oubliés à l'occasion du recensement sont plus susceptibles d'être classés dans les catégories "locataire" ou "ayant déménagé" (61% et 8%, respectivement) que les ménages dument dénombrés (46% et 0%, respectivement). Si l'on considère les ménages dans lesquels tous les adultes faisaient partie de la population active occupée en mars 1986, il y en a 12% de plus parmi les ménages oubliés que parmi les ménages dénombrés quoique, dans le cas des ménages oubliés, le nombre d'interviews soit trop peu important pour que cet écart soit significatif.

Ces résultats étayent la thèse selon laquelle les logements oubliés et les ménages oubliés dans des logements dument dénombrés présentent des caractéristiques qui réduisent leur visibilité au moment du recensement.

4.2 Ménages de la catégorie "non-appariement partiel"

Après les ménages de la catégorie "non-appariement intégral", voyons le cas des ménages classés dans la catégorie "non-appariement partiel". À ce stade de l'analyse, la comparaison porte sur 484 ménages ayant donné lieu à un non-appariement partiel et 331 ménages ayant donné lieu à un appariement intégral. Dans le cas de cette dernière catégorie de ménages, on a éliminé des 282 cas d'appariement intégral relevés dans l'échantillon de l'enquête sur les causes du sous-dénombrement les ménages comptant une seule personne puisque ceux-ci ne pouvaient pas donner lieu à un non-appariement partiel.

On a considéré deux séries de facteurs explicatifs. La première correspond aux caractéristiques du ménage que l'on pense pouvoir associer aux erreurs d'interprétation des définitions, présentées plus haut comme des erreurs dues au fait que le concept de membre du ménage n'est pas toujours interprété de la même façon par le Bureau of the Census et par les recenseurs. La deuxième série d'indicateurs regroupent les facteurs qui sont associés à la non-déclaration délibérée de certaines personnes dans un ménage.

Erreurs d'interprétation des définitions

Les indicateurs propres aux erreurs d'interprétation des définitions sont notamment la taille et la composition du ménage, la capacité de parler l'anglais, le niveau de scolarité du répondant et le statut du questionnaire par rapport au suivi consécutif à la vérification. Les ménages de grande taille, ceux qui englobent des personnes appartenées à des degrés plutôt éloignés et des personnes non appartenées à la personne responsable du ménage, ceux dont les membres parlent à la maison une langue autre que l'anglais, ceux où le répondant a un faible niveau de scolarité et ceux qui n'ont pas été désignés pour faire l'objet d'un suivi à l'issue de la vérification étaient davantage susceptibles de se signaler par des erreurs d'interprétation des définitions. La figure 4 corrobore ces hypothèses. On constate en effet que les ménages de la catégorie "non-appariement partiel" (trait léger) sont nettement de plus grande taille que ceux qui ont été classés dans la catégorie "appariement intégral" (trait foncé): 45% des ménages de la première catégorie par opposition à 19% seulement de ceux de la seconde catégorie comptaient six membres ou plus. Par contre, 40% des ménages classés dans la catégorie "non-appariement partiel" comprennent seulement la famille nucléaire de la personne responsable du ménage, le pourcentage étant de 72% dans le cas des ménages de la catégorie "appariement intégral". On observe une légère tendance, quoique significative, à l'effet que les ménages de la catégorie "non-appariement partiel" soient moins susceptibles d'avoir été désignés pour faire l'objet d'un suivi à l'issue de la vérification. Les ménages classés dans la catégorie "non-appariement partiel" (83%) sont davantage susceptibles de parler à la maison une langue autre que l'anglais

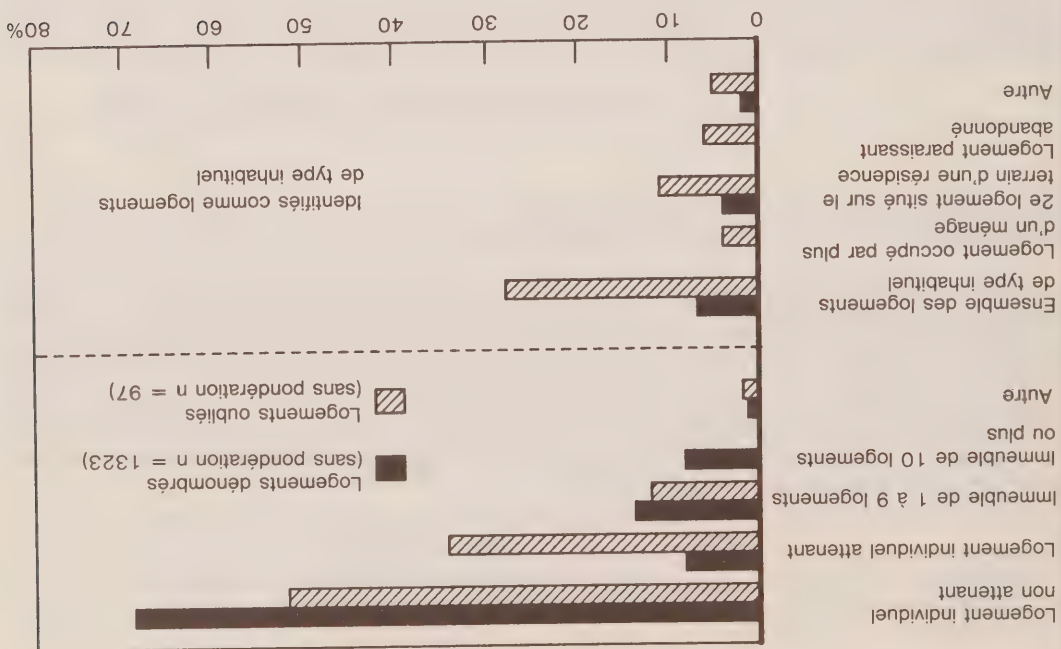


Figure 2. Caractéristiques matérielles des logements dûment dénombrés et des logements oubliés (pourcentages pondérés).

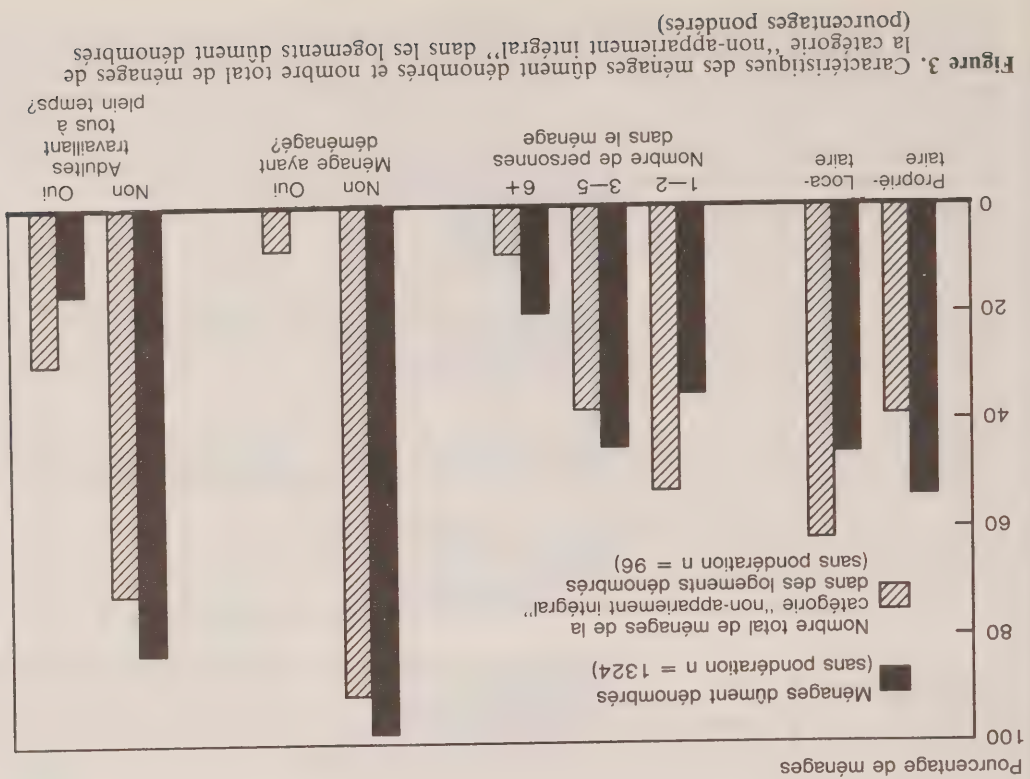


Tableau 2

Etat d'avancement au recensement des ménages de l'enquête postcensitaire
classés dans la catégorie "non-appartenance intégral", selon le statut
à l'égard du suivi des cas des non-réponse: nombre de logements^a

Etat d'avancement du logement au recensement	Envoyé pour le suivi des cas de non-réponse?	
	Non	Oui
Logement oublié sur les listes d'adresses utilisées pour le recensement	97	0
Logement inclus dans les listes d'adresses utilisées pour le recensement	4	92
Occupé, directement accepté ^b	1	6
Occupé, plaqué ^c	2	35
Inoccupé, directement accepté	1	34
Inoccupé, plaqué	0	17
Ensemble des logements	101	92
		193

Nota: a Nombre non pondéré.

b Directement accepté: le nombre de personnes donné par le programme FOSDIC est accepté tel quel.

c Plaqué: le nombre relevé sur le terrain est accepté à la place du nombre donné par le programme FOSDIC.

l'ensemble des ménages qui n'ont pu être apparés et dont le logement avait été répertorié comme étant occupé lors du recensement, 84% ont été dénombrés à l'occasion d'interviews subterfuges. La figure 2 donne une comparaison de certaines des caractéristiques matérielles des logements oubliés sur les listes d'adresses pour le recensement (trait léger) et des logements n'ont pas été oubliés (trait foncé). Le premier ensemble de traits correspond à des logements de type classique. Les maisons individuelles attenantes, comme les duplex, semblent avoir posé un problème sérieux au recensement d'essai de Los Angeles. En effet, 34% des logements oubliés appartiennent à cette catégorie qui regroupe seulement 8% des logements dûment dénombrés. Il y a moins de logements oubliés que de logements dûment dénombrés parmi les maisons individuelles non attenantes et les immeubles à appartements, ce qui laisse supposer que ces derniers types de logements sont plus faciles à repérer à l'occasion du recensement. Qu'ils aient pu ou non effectuer une interview, les intervieweurs de l'enquête sur les causes du sous-dénombrement devaient répertorier tout logement se classant dans une des catégories de logement de "type inhabituel" décrites sur la page couverture du questionnaire. La partie inférieure de la figure 2 indique que les intervieweurs ont identifié un pourcentage plus élevé (28%) de logements de type inhabituel parmi les logements qui ne figuraient pas sur les listes d'adresses utilisées pour le recensement que parmi les logements qui y figuraient (7%). Les types de logement qui ont posé des difficultés particulières sont les habitations qui semblent abandonnées et les logements secondaires situés sur le terrain d'une résidence. Les caractéristiques matérielles des logements ont donc une effet sur leur visibilité au moment de l'établissement des listes d'adresses pour le recensement. Quelle peut être la cause de l'oubli d'un ménage dans un logement qui a été dûment dénombré? Les ménages sont d'autant plus facilement oubliés qu'ils sont petits et mobiles. La figure 3 établit une comparaison entre les caractéristiques des ménages situés dans des logements dûment dénombrés qui n'ont pu être apparés et celles d'un groupe combinant des ménages classés dans les catégories "appartenance intégral" et "non-appartenance partiel", c'est-à-dire de ménages ayant été dénombrés. Les ménages oubliés lors du recensement d'essai (trait léger) étaient en

La dernière étape de l'analyse consiste à comparer les caractéristiques de quatre types de personnes: les personnes qui ont pu être *appariées* dans les ménages des catégories "non-appariement partiel" et "non-appariement partiel", et celles qui n'ont *pas pu être appariées* dans les ménages des catégories "non-appariement partiel" et "non-appariement intégral". La comparaison porte notamment sur l'âge, le sexe, le niveau de scolarité, le lien avec la personne responsable du ménage et le statut de citoyen(ne).

Les pourcentages résultant de la comparaison variables deux à deux sont basés sur des données pondérées afin de pallier l'effet des plans d'échantillonnage de l'enquête postcensitaire et de l'enquête sur les causes du sous-dénombrement quoique, pour les test relatifs aux écarts entre ces pourcentages, on utilise des chiffres non pondérés. On a également utilisé des données non pondérées pour estimer les paramètres des modèles log-linéaires. Pour évaluer l'incidence du plan d'échantillonnage de l'enquête postcensitaire sur les estimations appliquées dans les modèles finals, on a ajouté les interactions entre variables deux à deux, dont la variable de stratification de l'enquête postcensitaire. Cet ajustement n'a guère modifié les résultats, en conséquence, les estimations citées dans le présent article ne tiennent pas compte de la variable de stratification. Toutefois, du fait que le deuxième degré d'échantillonnage pour l'enquête postcensitaire nécessite l'échantillonnage en grappes des ménages en fonction des ilots définis pour le recensement, les erreurs-types obtenues risquent de sous-estimer la véritable erreur due à l'échantillonnage: elles sont données à titre indicatif de l'importance des paramètres.

4. RÉSULTATS

4.1 Ménages de la catégorie "non-appariement intégral"

Le tableau 2 indique l'état final attribué dans le cadre du recensement aux ménages classés, lors de l'enquête postcensitaire, dans la catégorie "non-appariement intégral" qui ont ou non été visés par le suivi des cas de non-réponse. Sur les 193 cas en question, 97 (c'est-à-dire 50%) ne figuraient pas sur les listes d'adresses utilisées pour le recensement. Autrement dit, l'oubli de logements semblerait être la raison pour laquelle on n'a pu retracer personne dans les fichiers du recensement pour ces ménages au moment de l'enquête postcensitaire.

Les 96 cas restants ne figuraient pas sur les listes d'adresses utilisées pour le recensement. Pour quelle raison ces ménages ont-ils été oubliés? L'explication tient probablement à ce que, dans la plupart des cas, il y a eu une interview subterfuge, c'est-à-dire que le propriétaire ou un voisin a donné uniquement une estimation du nombre total de personnes dans le ménage et aucun détail sur les membres du ménage. Cette supposition est étayée par le fait que, sur les 44 cas où le logement a été classé comme étant occupé au moment du recensement, pour 37 logements le nombre d'occupants a été "plaque". En d'autres mots, le chiffre définitif accepté pour ces ménages n'a pas été obtenu de la manière habituelle, c'est-à-dire par le décompte des occupants effectué automatiquement au moyen du programme FOSDIC (Film Optical Sensing Device for Input to Computers), mais on s'est contenté de "plaquer" le nombre de personnes déclaré sur le questionnaire du ménage en question au moment de la collecte des données sur le terrain. De telles circonstances sont très probablement une indication que le ménage a fait l'objet d'une interview subterfuge.

Autrement dit, la majorité de ces 44 ménages n'ont pas véritablement été oubliés lors du recensement même si, au moment d'apparier les données de l'enquête postcensitaire, on n'a pas trouvé d'enregistrements correspondants dans le fichier du recensement. Il est tenu compte de ces cas dans la méthode d'estimation basée sur les deux systèmes. Toutefois, il demeure que ces ménages n'ont pas été dénombrés de façon directe.

En résumé, 50% des ménages de l'enquête postcensitaire n'ayant pu être appariés correspondent à des logements qui ont été complètement oubliés. Pour ce qui est des ménages oubliés dans des logements dûment dénombrés, dans 54% des cas il s'agit de logements classés inoccupés, probablement par erreur, et dans 46% des cas, de logements qui étaient occupés en fait. Sur

Tableau 1

Nombre de ménages dans l'échantillon de l'enquête postcensitaire et dans celui de l'enquête sur les causes du sous-dénombrement et nombre d'interviews terminées, selon le type de ménage

Type de ménage	Enquête post-censitaire	Enquête sur les causes du sous-dénombrement	
		Echantillon	Interviews terminées
Appariement intégral	4,871	489	382
Non-appariement partiel	738	738	484
Non-appariement intégral	205	193	100
Ensemble des ménages	5,814	1,420	966

3.2 Plan d'analyse

L'analyse comporte plusieurs parties. Elle porte en premier lieu sur les ménages visés par l'enquête postcensitaire qui ont été classés dans la catégorie "non-appariement intégral". La comparaison est faite en deux étapes: (1) entre les logements oubliés et les logements dénombrés et (2) entre les ménages oubliés dans des logements dénombrés et les ménages dénombrés. On s'attendait à ce que les logements oubliés de logements groupés et de logements d'un type inhabituel ou situés dans un endroit inattendu que les logements dénombrés. On s'attendait par ailleurs à ce que les ménages oubliés dans des logements dénombrés soient de taille plus petite, composés d'adultes souvent absents et plus portés à déménager que les ménages dûment dénombrés. La plupart des variables explicatives se rapportant aux logements et aux ménages oubliés ont été tirées du fichier de contrôle des adresses du recensement ou du fichier des ménages appariés de l'enquête postcensitaire (qui contient les données du recensement et les données tirées de l'enquête postcensitaire pour chacun des ménages appariés) et sont donc disponibles pour les 193 ménages de l'échantillon classés dans la catégorie "non-appariement intégral".

La seconde étape consiste à comparer les ménages classés dans la catégorie "non-appariement partiel" avec ceux de la catégorie "appariement intégral" afin de mettre en évidence les facteurs à l'origine de l'oubli de certaines personnes à l'intérieur du ménage. Il existe deux séries de facteurs explicatifs: d'une part, les erreurs dues à un manque d'attention ou à une mauvaise interprétation des définitions et, d'autre part, les raisons qui ont entraîné une non-déclaration délibérée de certaines personnes dans le ménage. Les indicateurs typiques des erreurs dues à une mauvaise interprétation des définitions sont la grande taille du ménage ou sa composition complexe, une méconnaissance de l'anglais ou un très faible niveau de scolarité. Les indicateurs propres à la non-déclaration délibérée de certaines personnes dans le ménage sont la présence d'immigrants récents ou de bénéficiaires du bien-être, un logement surpeuplé ou l'incréduité en ce qui a trait au caractère confidentiel des données. On a posé comme hypothèse que les ménages classés dans la catégorie "non-appariement partiel" seraient davantage visés par les indicateurs d'une mauvaise interprétation des définitions et ceux de non-déclaration délibérée de certaines personnes dans le ménage que les ménages de la catégorie "appariement intégral".

L'analyse porte d'abord sur les liens entre deux variables, soit chacun des facteurs explicatifs et l'oubli de certaines personnes dans des ménages de la catégorie "non-appariement partiel", et ensuite sur les liens entre de multiples variables. Pour nombre des indicateurs étudiés, les données sont tirées de l'enquête sur les causes du sous-dénombrement; par conséquent, seules les données fournies par les ménages interviewés sont utilisées.

d'habitation les plus courants sont les maisons individuelles (73%) et les petits immeubles à appartements (15%). La moitié (51%) des logements occupés sont habités par le propriétaire, alors que le pourcentage correspondant pour l'ensemble du pays est de près des deux tiers (65%) (U. S. Bureau of the Census 1987a, page 106, tableau 18; U. S. Bureau of the Census 1987c, page 712, tableau 1285).

Les données analysées sont tirées des résultats du recensement d'essai de 1986 tenu à Los Angeles, de l'enquête postcensitaire effectuée en vue de mesurer la couverture du recensement d'essai et du suivi spécial réalisé à la suite de l'enquête postcensitaire sous le titre d'enquête sur les causes du sous-dénombrement. Un total de 109,900 logements a été dénombré lors du recensement de 1986 qui avait été conçu principalement en vue de mettre à l'essai les procédures prévues pour le recensement de 1990.

L'enquête postcensitaire, menée en juillet 1986, est précisément une des opérations mises à l'essai. Son but était de repérer les cas où des personnes avaient été oubliées ou mal dénombrées (Diffendal 1988). Pour ce faire, on a procédé à un appariement des données de l'enquête postcensitaire et des données du recensement. Lorsque, pour une personne dénombrée lors de l'enquête postcensitaire, on trouvait un enregistrement correspondant dans la base du recensement, on considérait qu'il y avait "appariement"; si on ne trouvait aucun enregistrement correspondant, on avait un cas de "non-appariement".

Les ménages couverts par l'enquête postcensitaire ont été répartis en trois catégories selon que la totalité, quelques-uns ou aucun des membres du ménage ont pu être appariés. La catégorie "appariement intégral" regroupe les ménages dont tous les membres dénombrés à l'occasion de l'enquête postcensitaire ont été retracés dans les données du recensement. La catégorie "non-appariement partiel" correspond aux ménages qui comptent un membre au moins n'ayant pu être retracé dans les données du recensement et un membre au moins ayant pu l'être. Enfin, la catégorie "non-appariement intégral" réunit strictement les personnes qui n'ont pu être appariées à aucun dossier du recensement.

On a établi cette distinction entre les trois types de ménages afin de pouvoir étudier les problèmes particuliers aux cas d'oubli du logement, d'oubli de tout un ménage dans un logement dûment dénombré et d'oubli de certaines personnes dans un ménage partiellement dénombré. Les ménages ayant donné lieu à un appariement intégral sont pris en considération à titre de référence, pour représenter les ménages convenablement dénombrés au recensement.

Un suivi spécial, l'enquête sur les causes du sous-dénombrement, a été effectué en novembre 1987 en vue de recueillir des renseignements complémentaires permettant de comparer ces trois types de ménage. Cette enquête a permis d'avoir des renseignements sur les caractéristiques de recensement pour les personnes n'ayant pu être appariées ainsi qu'un certain nombre de données concernant les ménages et les logements qui n'étaient pas disponibles dans les fichiers du recensement ni dans ceux de l'enquête postcensitaire.

Tous les ménages classés dans la catégorie "appariement partiel" et presque tous les ménages de la catégorie "non-appariement intégral" ont été sélectionnés pour une nouvelle interview. Huit ménages de la catégorie "non-appariement intégral" ont dû être éliminés de l'échantillon parce qu'il manquait plusieurs éléments pour pouvoir réaliser une interview. Un échantillon restreint a été tiré dans la catégorie "appariement intégral" afin de réduire les coûts du suivi.

La colonne située à l'extrême droite du tableau 1 donne la répartition, selon le type de ménage, des 966 ménages ayant répondu à une interview complète dans le cadre de l'enquête sur les causes du sous-dénombrement. Ce tableau donne également les chiffres non pondérés correspondant aux 5,814 ménages dénombrés lors de l'enquête postcensitaire et aux 1,420 ménages de l'échantillon de l'enquête sur les causes du sous-dénombrement. Le taux de réponse global pour le suivi est de 68%, ce qui représente un succès remarquable du point de vue de la localisation des ménages dans une région urbaine à population instable compte tenu de ce que l'intervalle entre l'enquête postcensitaire et le suivi était de seize mois.

Une autre opération importante du recensement est l'information du public. Les programmes de promotion du recensement sont conçus en vue d'inciter les recensés à retourner leur questionnaire par la poste et de réduire la non-déclaration délimitée en informant la population des utilisations des données du recensement, de la nécessité de renvoyer des questionnaires complets et du caractère confidentiel des enregistrements du recensement. Il n'a pas été établi dans quelle mesure ces programmes permettent de réduire les oublis de personnes à l'intérieur du ménage.

2.2 Contexte social

À chaque étape du dénombrement, les procédures de collecte des données se déroulent dans un contexte social qui présentent plusieurs caractéristiques susceptibles de nuire au bon déroulement du dénombrement. Ce peut être, par exemple, le refus de déclarer certains ou tous les membres du ménage, le fait que le répondant n'est pas en mesure de remplir le questionnaire en conformité avec les définitions établies ou encore le manque de "visibilité sociale" des membres du ménage ou du logement que ceux-ci occupent. (Par "visibilité sociale", on entend la mesure dans laquelle les membres du ménage ou le logement possèdent des caractéristiques qui les rendent visibles aux personnes de l'extérieur.)

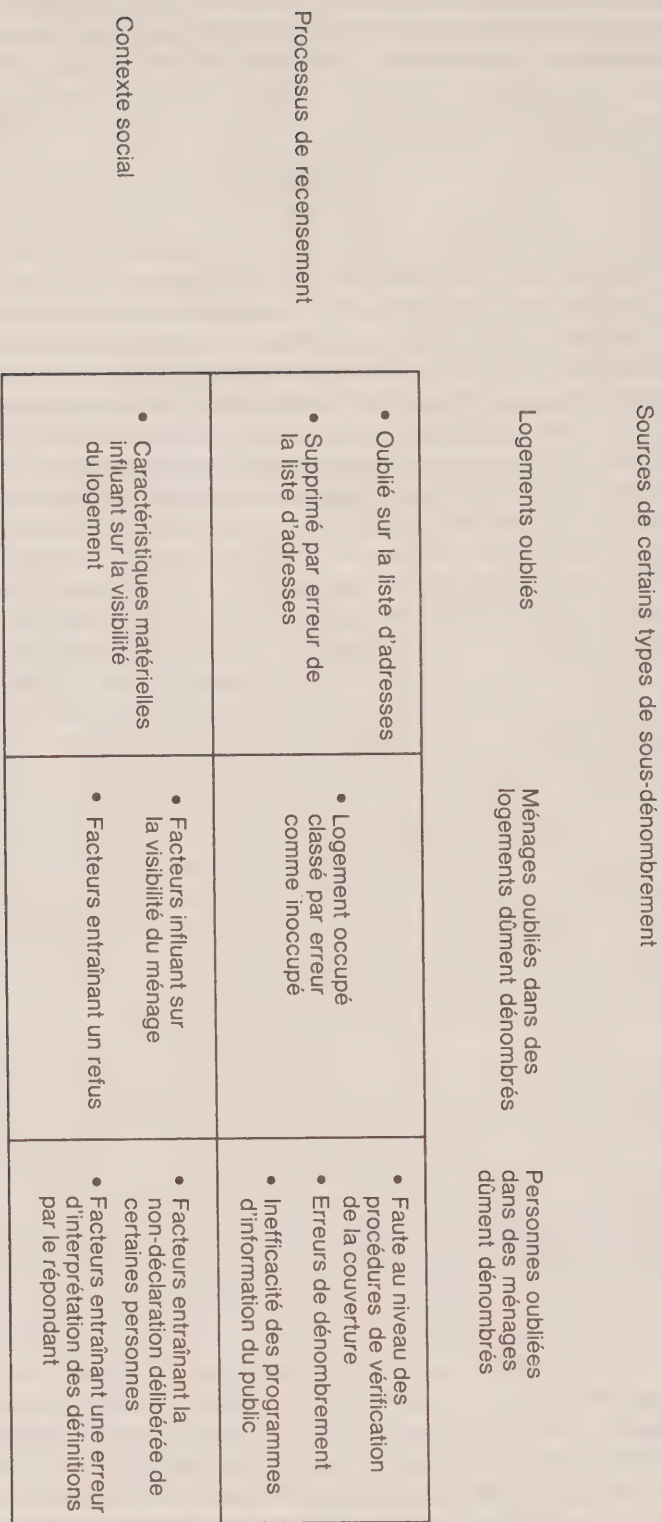
Les principaux facteurs du contexte social qui peuvent occasionner l'oubli d'un logement sont ceux qui nuisent à sa visibilité sociale. Certains types de logements sont plus faciles à trouver et sont davantage susceptibles de figurer sur les listes d'adresses commerciales que d'autres. Les causes d'oubli de ménages vivant dans des logements dûment dénombrés incluent notamment les facteurs nuisant à la visibilité des membres du ménage et le refus de répondre au questionnaire. Ces trois grandes sources de sous-dénombrement lié au contexte social sont, au même titre, susceptibles d'être à l'origine de l'oubli de personnes à l'intérieur du ménage: refus de répondre au questionnaire, mauvaise interprétation des définitions du recensement et manque de visibilité sociale des membres du ménage. La question de la participation des recensés peut être abordée sous l'angle des coûts et avantages que le recense perçoit pour retirer de sa participation (Dillman 1978). On a beaucoup débattu des coûts que la participation au recensement peuvent représenter pour les recensés. Certains peuvent redouter que le fait de déclarer la présence dans le ménage d'adultes de sexe masculin entraîne l'inadmissibilité aux programmes de bien-être, que le fait de déclarer la présence dans le ménage de personnes entrées illégalement au pays entraîne l'expulsion de celles-ci, que le fait de déclarer la présence de plus de personnes que le bail n'en autorise entraîne une intervention du propriétaire ou encore que la police soit mise au courant de l'endroit où se cachent des personnes qui ont transgressé la loi (Bailar et Martin 1987). De telles craintes peuvent susciter la non-participation lorsque la personne doute que le Bureau of the Census respecte son engagement à préserver le caractère confidentiel des données. Le problème est très différent dans le cas des erreurs d'interprétation des définitions. De telles erreurs peuvent se produire lorsque la situation des particuliers dans le ménage est complexe et que le répondant a du mal à comprendre et à appliquer les instructions relatives au dénombrement des membres du ménage et à la résidence (Hainer et coll. 1988). Après ce survol de quelques-unes des principales causes de sous-dénombrement, nous allons analyser leur incidence lors du recensement d'essai effectué en 1986 à Los Angeles.

3. MÉTHODES

3.1 Sources de données

La présente étude porte plus spécifiquement sur le sous-dénombrement au recensement d'essai effectué en mars 1986 dans la moitié septentrionale du comté de Los Angeles. L'univers observé est une sous-population à faible revenu et à dominante hispanique. Près des deux tiers (65%) des responsables des ménages dénombrés lors du recensement d'essai étaient d'origine espagnole et 13% d'origine asiatique. Dans cette partie de la ville de Los Angeles, les types

Figure 1. Modèle de recherche



Bien que les prétendues injustices mentionnées plus haut résultent d'une erreur de couverture *netie* (personnes oubliées moins personnes dénombrées par erreur), afin de pouvoir procéder à l'analyse, on a tenu compte uniquement des personnes oubliées. Le cas des personnes oubliées au recensement des États-Unis mérite un traitement prioritaire dans le calendrier de la recherche relative au recensement du fait que ces oubliés sont plus nombreux, qu'ils varient de façon plus systématique en fonction des caractéristiques socio-économiques et qu'ils sont plus controversés politiquement que les cas de personnes dénombrées par erreur.

L'article présente, en premier lieu, le système utilisé pour classer les causes du sous-dénombrement. Il expose ensuite les méthodes employées et les résultats obtenus. En conclusion, il résume les façons possibles d'améliorer la couverture.

2. MODÈLE DE RECHERCHE

Le modèle de recherche est illustré par la figure 1. Le sous-dénombrement y est présenté comme un problème qui se pose essentiellement au niveau du ménage plutôt qu'à celui de la personne. Cette spécification est fondée sur les sources mêmes du sous-dénombrement dans un recensement où l'on cherche à entrer en contact avec chaque ménage plutôt que chaque personne.

La marge supérieure de la figure 1 fait état de trois problèmes de sous-dénombrement au niveau du ménage: oubli d'un ménage au complet parce qu'un logement n'a pas été dénombré, oubli d'un ménage au complet dans un logement qui a été dénombré et oubli de certaines personnes dans un ménage dont les autres membres ont été dénombrés. Ces trois problèmes peuvent survenir à cause de la façon dont les opérations de dénombrement sont effectuées, à cause de la réaction de certains membres de la société dénombrée ou par suite de l'interaction du contexte social et des mesures opérationnelles. La section suivante porte uniquement sur les erreurs liées à l'utilisation de la méthode d'envoi et retour par la poste du questionnaire appliqué lors du recensement d'essai mené en 1986 auprès d'une sous-population à faible revenu et à dominante hispanique.

2.1 Mise en oeuvre des opérations de recensement

Des difficultés opérationnelles survenues au cours du dénombrement peuvent entraîner l'oubli de certains logements, de ménages à l'intérieur de logements dument dénombrés ou de personnes à l'intérieur de ménages dument dénombrés. Des logements occupés peuvent être oubliés parce que, à aucune étape, ils n'ont été rajoutés sur les listes d'adresses ou parce qu'ils ont été éliminés par erreur de la liste sur laquelle ils figuraient (U. S. General Accounting Office 1980). Lorsque le logement est bien répertorié sur la liste d'adresses, tous ses occupants peuvent néanmoins avoir été oubliés si le logement a été classé par erreur comme logement inoccupé lors du suivi des cas de non-réponse (U. S. Bureau of the Census 1987b; Eriksen 1983).

Dans le cas des questionnaires postaux remplis et retournés par le ménage, il y a peu de moyens de repérer les personnes oubliées. Les procédures visant à améliorer la couverture à l'intérieur des ménages comprennent une question par laquelle le répondant doit dire s'il a eu des doutes quant à la nécessité d'inclure une personne en particulier dans le questionnaire et une vérification manuelle de la cohérence des données qui consiste à comparer la liste des membres du ménage que le répondant a dressée au début du questionnaire et le nombre de personnes pour lesquelles des renseignements sont fournis dans la suite du questionnaire (U. S. Bureau of the Census 1987b; Edson 1987). Ces procédures peuvent occasionner des oublis à l'intérieur du ménage lorsqu'elles ne produisent pas l'effet voulu par suite d'une mauvaise administration du suivi. De même, les erreurs commises par les recenseurs lors du suivi des questionnaires non retournés par la poste peuvent empêcher que des personnes oubliées soient ajoutées.

Sources du sous-dénombrement lors du recensement: Résultats du recensement d'essai de 1986 à Los Angeles

DAVID J. FEIN et KIRSTEN K. WEST¹

RÉSUMÉ

Le présent article expose les résultats d'une étude des causes du sous-dénombrement à l'occasion du recensement d'une région urbaine à dominante hispanique particulièrement difficile à dénombrer. L'étude propose un cadre d'organisation des causes du sous-dénombrement et tente d'expliquer ces celles-ci à partir de diverses hypothèses. L'approche adoptée est unique dans le sens qu'elle vise à quantifier les causes de sous-dénombrement et à isoler les problèmes exceptionnellement importants en incluant une analyse statistique des autres problèmes.

MOTS CLÉS: Recensement; sous-dénombrement; amélioration de la couverture; enquête postcensitaire.

1. INTRODUCTION

Au cours des deux dernières décennies à peu près, le besoin de mieux comprendre les causes du sous-dénombrement observé au recensement des États-Unis est devenu de plus en plus impératif. Par suite de l'importance grandissante du recensement comme outil pour gouverner la nation, gérer les affaires et contrôler l'évolution de la société (Citro et Cohen 1985; Clogg et coll. 1986), le public se montre de plus en plus soucieux de la qualité des données du recensement. Cette préoccupation vient en grande partie de ce qu'il apparaît, à juste titre, que le sous-dénombrement net qui se produit à l'occasion du recensement a un effet disproportionné sur les membres moins favorisés de la société (Citro et Cohen 1985, chap. 5; Erickson 1983). Les représentants des sous-populations défavorisées estiment que le sous-dénombrement entraîne, pour ces dernières, une perte importante au chapitre de la part des fonds publics qui leur revient ainsi qu'une sous-représentation politique (Choldin 1987).

En admettant qu'il soit possible de trouver une méthode acceptable, une des solutions à ce problème consisterait à corriger les chiffres du recensement en fonction du biais occasionné par le sous-dénombrement. Toutefois, à l'automne 1987, le Department of Commerce a décidé de ne pas procéder à un tel ajustement pour le recensement de 1990 mais de chercher à obtenir un dénombrement plus précis (Ortner 1987).

La décision d'améliorer la couverture du recensement conduit au besoin de comprendre mieux que jamais auparavant les causes du sous-dénombrement qui survient au recensement. Un certain nombre de programmes spéciaux d'amélioration de la couverture ont été mis sur pied lors du recensement de 1980 et ceux-ci peuvent expliquer le fait que l'on ait enregistré les plus faibles taux globaux d'erreur de couverture nette jamais vus. Malgré cela, d'importants écarts de couverture de portée socio-économique ont persisté. Le Bureau of the Census a donc entrepris un vaste programme de recherche en vue d'identifier les causes du sous-dénombrement, en mettant l'accent principalement sur les sous-populations particulièrement difficiles à dénombrer.

Le présent article expose les résultats d'une étude des causes du sous-dénombrement survenu lors du recensement dans un quartier de Los Angeles à dominante hispanique. L'approche adoptée est unique dans le sens qu'elle vise à quantifier les causes de sous-dénombrement et à isoler les problèmes d'importance exceptionnelle en contrôlant statistiquement les autres problèmes.

¹ David J. Fein et Kirsten K. West, Undercount Research Staff, Statistical Research Division, Bureau of the Census, Washington, D. C. 20233. Les opinions exprimées dans cet article sont celles des auteurs et n'engagent nullement le Bureau of the Census.

- 236 Rubin, Schaffer, et Schenker: Méthodes d'imputation de valeurs manquantes
- RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.
- RUBIN, D.B., SCHAFER, J.L., et SCHENKER, N. (1988). Imputation strategies for estimating the undercount. *Proceedings of the Fourth Annual Research Conference*, United States Bureau of the Census, 151-159.
- SCHENKER, N. (1988). Traitement des données manquantes dans l'estimation de la couverture: le test des opérations de redressement de 1986, *Techniques d'enquête*, 14, 93-104.
- SCHENKER, N. (1989). The use of imputed probabilities for missing binary data. *Proceedings of the Fifth Annual Research Conference*, United States Bureau of the Census (à paraître).
- WOLTER, K.M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, 81, 338-346.

REMERCIEMENTS

Cet article est le compte rendu de recherches qui ont été faites principalement lorsque Nathaniel Schenker travaillait à la Statistical Research Division du U.S. Bureau of the Census, Washington, DC 20233, E.-U. Les opinions qui y sont exprimées sont celles des auteurs et ne reflètent pas nécessairement la position du U.S. Bureau of the Census. Cette étude a été rendue possible en partie grâce aux conventions sur la statistique n° 87-07 et 88-02 entre le U.S. Bureau of the Census et Harvard University, et en partie par la U.S. National Science Foundation sous subvention SES-88-05433, et est une version révisée et améliorée de Rubin, Schaffer et Schenker (1988). Les auteurs tiennent à remercier les deux arbitres pour leurs commentaires très précieux.

BIBLIOGRAPHIE

BAKER, S.G., et LAIRD, N.M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association*, 83, 62-69.

BISHOP, Y.M.M., FIENBERG, S.E., et HOLLAND, P.W. (1975). *Discrete Multivariate Analysis*. Cambridge: MIT Press.

DEMPTSTER, A.P., LAIRD, N.M., et RUBIN, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39, 1-38.

DIFFENDAL, G. (1988). Test des opérations de redressement de 1986 dans Central Los Angeles County. *Techniques d'enquête*, 14, 71-92.

FAY, R.E. (1986). Causal models for patterns of nonresponse. *Journal of the American Statistical Association*, 81, 354-365.

FAY, R.E., PASSEL, J.S., et ROBINSON, J.G. (1988). *The Coverage of Population in the 1980 Census*. 1980 Census of Population and Housing Evaluation and Research Report PHC80-E4, Washington: U.S. Government Printing Office.

FREEDMAN, D.A., et NAVIDI, W.C. (1986). Regression models for adjusting the 1980 Census. *Statistical Science*, 1, 3-39.

HOGAN, H., et WOLTER, K. (1988). Mesure de l'erreur dans une enquête postcensitaire. *Techniques d'enquête*, 14, 105-124.

KROTKI, K.J. (1978). *Developments in Dual System Estimation of Population Size and Growth*. Edmonton: The University of Alberta Press.

LAIRD, N.M. (1978). Empirical Bayes methods for two-way contingency tables. *Biometrika*, 65, 1, 581-590.

LEONARD, T. (1975). Bayesian estimation methods for two-way contingency tables. *Journal of the Royal Statistical Society, Series B*, 37, 23-37.

LITTLE, R.J.A. (1985). Nonresponse adjustments in longitudinal surveys: models for categorical data. *Bulletin of the International Statistical Institute*, 15, 1-15.

LITTLE, R.J.A. (1986). Missing data in Census Bureau surveys. *Proceedings of the Second Annual Research Conference*, United States Bureau of the Census, 442-454.

LITTLE, R.J.A., et RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

MARKS, E.S., SELTZER, W., et KROTKI, K.J. (1974). *Population Growth Estimation*. New York: The Population Council.

RUBIN, D.B. (1974). Characterizing the estimation of parameters in incomplete-data problems. *Journal of the American Statistical Association*, 69, 467-474.

RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 3, 581-592.

calculer. Une méthode consiste essentiellement à tirer plusieurs valeurs de Θ de cette distribution *a posteriori* sans déterminer de façon explicite le mode *a posteriori* ou la moyenne; les valeurs de Θ ainsi prélevées peuvent servir à imputer par multiplication les données manquantes.

À la question de la mesure de la variabilité se rattache celle de l'efficacité dans l'échantillon-nage répété. Bien que nous croyions que notre méthode bayésienne convient parfaitement au problème, il importe, pour assurer une application générale, d'évaluer les caractéristiques d'application de cette méthode dans toutes les circonstances où elle pourrait être appliquée couramment. Par exemple, quelle est son efficacité dans les cas concrets lorsque la non-réponse est aléatoire et que l'analyste de données l'ignore?

Ces sujets feront l'objet d'importants travaux de recherche.

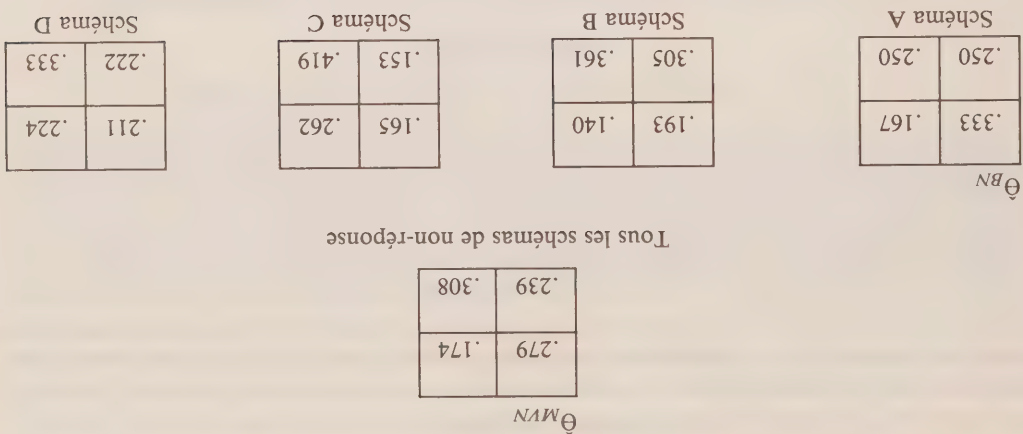
4.2 Nécessité d'évaluer l'opportunité relative des modèles

Étant donné le modèle saturé pour (X, Y, R) avec le paramètre Θ , n'importe quelle méthode d'imputation, si incohérente soit-elle, peut être considérée comme la méthode appropriée selon un modèle quelconque. Prenons, par exemple, une méthode d'imputation qui utilise Θ_{CC} comme tableau de référence pour tous les schémas de non-réponse. Cette méthode énonce des distributions conditionnelles pour les données manquantes, étant donné les observations et R , au sujet desquelles les valeurs observées ne nous apprennent rien. Par conséquent, le fait de combiner ces distributions à des distributions qui peuvent être estimées (les distributions de R et des observations) suppose un estimateur de Θ qui maximise la vraisemblance selon le modèle saturé! Ce n'est pas une solution très sensée puisqu'elle correspond à l'unique BMV dans un modèle où toutes les sortes de distributions conditionnelles, étant donné divers schémas de non-réponse R , sont égales aux distributions conditionnelles, étant donné $R = (1, 1, \dots, 1)$; toutefois, si nous considérons uniquement la fonction de vraisemblance, il n'y a aucune raison de préférer un autre estimateur du maximum de vraisemblance à celui-là.

Même les méthodes d'imputation plus bizarres, comme l'imputation de zéros à la place de toutes les valeurs manquantes, correspondent à des modèles particuliers où les estimateurs de Θ sont des EMV en vertu du modèle saturé; cependant, ces méthodes sont contraires au bon sens. Toute tentative sérieuse d'imputation repose sur l'idée que deux personnes pour lesquelles les valeurs des caractéristiques observées et les schémas de non-réponse sont les mêmes ne sont pas si loin l'une de l'autre en ce qui concerne les caractéristiques qui ont été observées pour l'une mais non pour l'autre. Notre méthode BN formalise cette notion de rapprochement en définissant un modèle de tableau de contingence avec de faibles interactions de degré supérieur.

En conclusion, le choix d'une méthode d'imputation parmi d'autres ne peut être dicté uniquement par des principes qui tiennent du maximum de vraisemblance; il faut aussi examiner la pertinence des conditions préalables fondamentales. Cela ne pose pas réellement de problème sérieux; la méthodologie statistique a toujours recommandé l'utilisation de modèles simples ou uniformes dans les cas où des modèles moins uniformes ajustent tout aussi bien les données. Supposons que l'on veuille ajuster des droites ou des courbes polynomiales à travers une série de points; pour des motifs scientifiques, on préférera les modèles simples aux modèles complexes — il en est de même pour l'imputation. Nous croyons que le modèle qui est à la base de la méthode BN (et qui est défini par les équations (2) et (3)) conviendra dans beaucoup de problèmes, tout comme la régression linéaire est un outil acceptable dans de nombreux problèmes.

Figure 2. Tableaux de référence pour l'imputation



Observations

Valeurs imputées θ_{CC}

Valeurs imputées θ_{MVN}

Valeurs imputées θ_{BN}

$X = 1$	100	50
$X = 0$	75	75

$Y = 1 \quad Y = 0$

Tableau A

$X = 1$	30
$X = 0$	60

$Y = 1 \quad Y = 0$

Tableau B

28	60
----	----

$Y = 1 \quad Y = 0$

Tableau C

12

Tableau D

3	4
3	2

2.86	3.35
3.69	2.09

2.68	2.54
4.09	2.70

12	16
36	24

12.9	15.1
38.3	21.7

13.5	14.5
36.9	23.1

30	20
30	10

26.2	18.5
33.8	11.5

27.5	17.4
32.5	12.6

départ, l'algorithme converge autour de quatre cycles environ. L'estimateur Θ_{NB} (méthode bayésienne «non neutre») est déterminé au moyen d'une distribution *a priori* où $\sigma^2 = 10$ et $\tau = 3$. Cela signifie que les termes du premier degré sont distribués *a priori* selon une loi normale de moyenne nulle et de variance 10, de sorte qu'il y a 95% de chances que le logit pour chaque effet principal se situe dans l'intervalle $(-4\sqrt{10}, +4\sqrt{10})$. La variance est de $10/3$ pour les termes du second degré, de $10/9$ pour les termes du troisième degré et de $10/27$ pour les termes du quatrième degré; ces chiffres indiquent que les termes de degré supérieur tendent modérément vers l'origine. (Il a été difficile de déterminer Θ_{NB} pour diverses valeurs de σ^2 et de τ à cause de l'instabilité numérique de l'algorithme de maximisation particulier appliqué à chaque phase M.) Les valeurs de Θ_{IML} et de Θ_{NB} sont présentées à la figure 2. Les valeurs imputées espérées pour ces modèles sont présentées à la figure 3 et à des fins de comparaison les valeurs imputées espérées pour le modèle Θ_{CC} y sont aussi présentées.

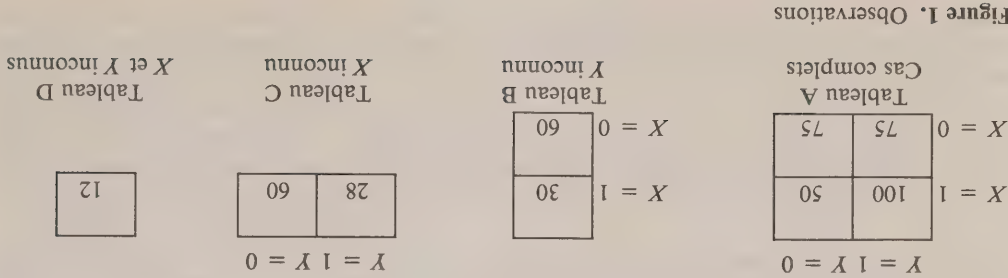
On peut voir très clairement les différences entre les méthodes d'imputation en comparant les valeurs imputées espérées pour le tableau D. La méthode d'imputation fondée sur Θ_{CC} reproduit simplement les proportions observées dans le tableau A. La méthode d'imputation fondée sur Θ_{IML} diffère de l'imputation fondée sur Θ_{CC} car les tableaux B et C, de même que le tableau A, contribuent à l'estimation de Θ et de ce fait à l'imputation pour le tableau D. La méthode d'imputation fondée sur Θ_{NB} diffère fondamentalement de l'imputation fondée sur Θ_{CC} ou sur Θ_{IML} en ce sens qu'elle suppose que la non-réponse est informative. Partant du tableau B, elle suppose que l'absence de X est liée à $X = 0$. Partant du tableau C, elle suppose que l'absence de X est liée à $X = 0$. On combine ensuite ces renseignements d'une manière judicieuse pour en conclure qu'une proportion relativement plus forte des personnes pour lesquelles il manque X et Y se trouvent dans la catégorie ($X = 0, Y = 0$).

4. ANALYSE

Il est clair que nos recherches en sont à l'état embryonnaire. Nous croyons néanmoins qu'elles aboutiront à des applications importantes relativement à l'estimation du sous-dénombrement au moyen d'une EF et, de façon plus générale, à la modélisation de tableaux de contingence lorsqu'il manque certaines données. Nous allons conclure cet article par deux commentaires brefs: le premier sur la nécessité de poursuivre les recherches sur ce sujet et le second, sur la nécessité d'évaluer l'opportunité relative des modèles lorsqu'on élabore une méthode d'imputation.

4.1 Poursuivre les recherches

Deux aspects doivent être approfondis avant que la méthode BN puisse être appliquée de façon générale. Premièrement, il faut axer la recherche sur les méthodes de calcul si l'on veut approfondir la distribution *a posteriori* en forme d'arrêt. Les paramètres susceptibles de remplacer le mode, comme la moyenne *a posteriori*, valent la peine d'être considérés. Par ailleurs, il conviendrait de calculer des mesures de variabilité et compte tenu de la forme non normale peu courante de la distribution *a posteriori*, ces mesures pourraient être difficiles à résumer ou à



où $\theta_{ijk...p}$ est la probabilité qu'une observation se trouve dans la cellule $ijk...p$, et les μ sont les interactions à un, à deux ou à trois degrés et les interactions de degré supérieur. Nous posons le groupe élémentaire de distributions *a priori* normales indépendantes

$$\begin{aligned} \mu_i &\sim N(0, \sigma^2) \\ \mu_j &\sim N(0, \sigma^2/\tau) \\ \mu_{jk} &\sim N(0, \sigma^2/\tau^2) \\ &\vdots \\ \mu_{ijk...p} &\sim N(0, \sigma^2/\tau^j), \end{aligned} \quad (3)$$

pour une valeur choisie de $\sigma^2 > 0$ et de $\tau > 1$. Cette distribution *a priori* tend à ramener les interactions de degré supérieur vers zéro et, de ce fait, favorise l'utilisation d'un modèle plus simple ou plus uniforme pour estimer θ . Nous croyons que cette méthode produira des valeurs estimées de θ qui ne sont pas trop différentes de θ_{IML} lorsque la non-réponse est réellement aléatoire, mais qu'elle sera plus robuste que la méthode MVN lorsque la non-réponse ne sera pas aléatoire. La seule occasion où la méthode MVN sera supérieure à la méthode proposée est lorsque la non-réponse est aléatoire et qu'il existe de fortes interactions de degré supérieur entre les X et les Y .

Leonard (1975) et Laird (1978) ont analysé, pour des données complètes, des modèles linéaires logarithmiques avec distributions *a priori* normales pour les termes μ ; dans notre cas, l'analyse se complique du fait que nous ne connaissons que certains effets marginaux du tableau octodimensionnel. Il est théoriquement simple de déterminer le mode *a posteriori* θ_{NB} BN selon ce modèle; on peut appliquer l'algorithme EM à la distribution *a posteriori* de θ , comme on l'applique à la fonction de vraisemblance. La phase E (espérance mathématique) ne change pas; la phase M (maximisation), toutefois, pose certains problèmes de calcul. La distribution *a posteriori* a presque la forme d'une arête dans l'espace de dimension supérieure; elle est très accentuée dans certaines directions mais presque aplatie dans d'autres. La matrice des dérivées secondes est quasi singulière le long de cette arête; par conséquent, la méthode de Newton-Raphson et les autres méthodes de maximisation fondées sur la méthode de la plus grande pente seront pas efficaces. La situation se complique lorsque σ^2 augmente puisque l'arête s'aplatit lorsque $\sigma^2 \rightarrow \infty$ et nous n'avons plus un mode unique. Des problèmes surgissent aussi lorsque le nombre d'observations augmente puisque la distribution *a posteriori* devient très accentuée dans certaines directions, ce qui amplifie sensiblement des sections de la matrice des dérivées secondes. Il faut poursuivre les recherches afin d'élaborer des méthodes efficaces pour calculer θ_{BN} .

3.4.4 Exemple numérique

Nous allons maintenant présenter un exemple numérique simple où nous comparons les résultats obtenus à l'aide des méthodes MVN et BN. Pour simplifier le problème, nous n'utilisons qu'une seule variable dichotomique X (qui prend les valeurs 0 ou 1) et la variable Y , qui désigne le code d'appariement.

S'il ne manquait aucune donnée, on pourrait classer toutes les données selon X et les présenter sommairement sous forme d'un tableau de contingence 2×2 . Or, comme il y a quatre schémas de non-réponse, les données sont présentées dans un tableau CC et trois tableaux supplémentaires (figure 1). L'estimateur θ_{CC} (méthode des cas complets) correspond simplement aux proportions observées dans le tableau A. L'estimateur θ_{IML} (méthode du maximum de vraisemblance «neutre») est calculé itérativement au moyen de l'algorithme EM; avec θ_{CC} comme valeur de

autre variable dont la valeur serait observée pour des personnes pour lesquelles $R_1 = 1$ et d'autres pour lesquelles $R_1 = 0$. Un exemple d'une quantité qui ne peut réellement pas être estimée à l'aide des données est $P(Y = 1 \mid X_1 = i, X_2 = j, X_3 = k, R_1 = R_2 = R_3 = 1, R_j = 0)$, mais cette expression ne correspond à aucun terme d'interaction dans la paramétrisation du modèle log-linéaire. (Lorsque nous parlons d'une quantité «qui ne peut réellement pas être estimée», nous voulons dire, comme Rubin (1974), que la distribution *a posteriori* du paramètre égale sa distribution *a priori*, quelle que soit cette distribution.)

Lorsqu'un ensemble de données présente un schéma de non-réponse complexe, il n'est pas facile d'y trouver une série de termes log-linéaires qui, une fois définis comme nuls, donneront un EMV unique pour θ . Le nombre minimum de termes qu'il faut définir comme nuls pour obtenir un EMV unique est $2^{JJK} - 1$, (la dimension de θ) moins le nombre de statistiques exhaustives observées. Même s'il est possible de trouver une telle série de termes, elle n'est habituellement pas unique et on est obligé, par conséquent, de décider de la série de termes qu'il faut exclure du modèle. Au lieu de tenter d'obtenir un EMV unique en soumettant le modèle log-linéaire à ce genre de conditions préalables, nous allons utiliser une méthode bayésienne qui prévoit l'usage d'une distribution *a priori*.

3.4.3 Méthode bayésienne «non neutre»

Dans le modèle bayésien, on exprime formellement les hypothèses préalables relatives aux paramètres par une distribution *a priori*. Dans le cas qui nous occupe, la combinaison d'une distribution *a priori* unimodale appropriée et de la fonction de vraisemblance des observations nous pouvons considérer le mode *a posteriori* θ_{NB} comme l'estimateur unique de θ . Cette méthode donne une distribution *a posteriori* qui peut produire un estimateur unique de θ ; par exemple, est intéressante du fait qu'elle permet automatiquement d'estimer avec précision les fonctions de θ au sujet desquelles il existe beaucoup de renseignements tout en puisant dans la distribution *a priori* des valeurs appropriées pour les quantités qu'il est strictement impossible d'estimer à l'aide des données. Si elle est appliquée convenablement, cette méthode donnera un modèle «non neutre» qui ajuste les données aussi bien que n'importe quel autre modèle — il maximise essentiellement la fonction de vraisemblance et est néanmoins aussi conforme que possible à l'idée que nous nous faisons de la nature du mécanisme de non-réponse telle qu'elle est exprimée dans la distribution *a priori*.

La méthodologie scientifique devrait nous faire choisir une distribution *a priori* qui implique une structure simple (c'est-à-dire faibles interactions de degré supérieur) plutôt qu'une structure complexe (c'est-à-dire fortes interactions de degré supérieur). Si nous choisissons une distribution *a priori* qui associe une probabilité *a priori* peu élevée (mais non nulle) à la présence d'interactions de degré supérieur dans le modèle log-linéaire, nous allons poser des hypothèses qui sont de même nature que celles de la méthode MVN — à savoir que les valeurs manquantes ne diffèrent pas absolument des valeurs observées correspondantes en ce qui regarde leur rapport avec d'autres variables observées — sauf que cela se fera de façon plus uniforme et plus systématique.

En adoptant la notation de Bishop, Fienberg et Holland (1975), considérons le modèle linéaire logarithmique saturé pour le tableau de contingence octodimensionnel pour R , X et Y ,

$$\log \theta_{ijk\dots p} = \mu + \mu_{1(i)} + \mu_{2(j)} + \dots + \mu_{8(p)} + \mu_{12(ij)} + \mu_{13(ik)} + \dots + \mu_{123\dots 8(ijk\dots p)}$$

(2)

aléatoire, l'utilisation de θ^{CC} à des fins d'imputation a pour effet d'introduire des biais dans les données. De plus, même lorsque la non-réponse est parfaitement aléatoire, θ^{CC} n'est pas

efficace car il n'utilise pas toutes les observations pour estimer θ .

En revanche, la méthode MVN utilise toutes les données, aussi bien celles du tableau CC que celles des tableaux supplémentaires, pour estimer θ . La valeur estimée θ^{MVN} est choisie pour maximiser la vraisemblance en ne tenant pas compte du mécanisme de non-réponse (Little et Rubin 1987, section 5.3). De façon générale, il n'existe pas d'expression en forme analytique pour θ^{MVN} ; cette valeur doit être calculée itérativement en utilisant, par exemple, l'algorithme

EM (Dempster, Laird et Rubin 1977; Little et Rubin 1987, section 9.3).

L'algorithme EM pour les tableaux de contingence est facile à appliquer et l'estimateur du maximum de vraisemblance θ^{MVN} qui en découle est efficace et converge selon l'hypothèse de «neutralité»; par conséquent, l'algorithme EM est intéressant tant du point de vue théorique que du point de vue du calcul en ce qui concerne la méthode MVN. Cependant, lorsque la non-réponse n'est pas aléatoire, la méthode MVN introduit habituellement des biais dans l'estimation. Comme il y a de bonnes raisons de croire que la non-réponse dans l'EP n'est pas aléatoire, nous proposons ci-dessous une nouvelle méthode d'estimation qui repose sur une hypothèse différente.

3.4.2 Modèle «non neutre» et non-unité de l'EMV

Lorsque la non-réponse n'est pas aléatoire, il n'est plus permis d'ignorer le mécanisme de non-réponse; le fait qu'il manque un élément de réponse nous renseigne sur la valeur de cet élément. Par conséquent, un modèle qui tient compte de cette dépendance doit inclure des variables qui indiquent si la valeur d'un élément a été observée ou si elle est manquante. Ainsi, un modèle «non neutre» permettra habituellement d'estimer un tableau de référence pour chaque schéma de non-réponse ou bien un tableau de référence élargi, notamment un tableau ayant le double des dimensions initiales (c.-à-d. une dimension additionnelle pour chaque indicateur de non-réponse).

Soit $R = (R_1, R_2, R_3, R_Y)$ des variables qui indiquent respectivement si X_1, X_2, X_3 , et Y ont été observées; par exemple, $R_1 = 1$ si X_1 est connue et $R_1 = 0$ si X_1 est inconnue. Considérons le tableau de contingence à huit dimensions formé en répartissant les personnes selon X, Y et R et définissons θ comme le tableau octodimensionnel des probabilités de cellules pour ce tableau élargi.

Chaque personne visée par l'enquête est incluse dans une cellule du tableau élargi mais, comme il manque des données, nous ne connaissons que certains effets marginaux du tableau. Comme les valeurs de R sont toutes connues, les effets marginaux qui concernent uniquement des indicateurs de non-réponse sont tous connus alors que des effets marginaux qui se rapportent à Y ou à une des variables X peuvent ne pas être connus. Par exemple, dans un tableau de contingence où $R_1 = R_2 = R_3 = 1$ et $R_Y = 0$, les personnes peuvent être classées selon X_1, X_2 , et X_3 , mais non selon Y ; par conséquent, on ne peut connaître que les effets marginaux obtenus en faisant la somme par rapport à Y .

Le nombre de paramètres inclus dans le modèle saturé pour ce tableau est $2^5 IJK - 1$, ce qui dépasse le nombre de statistiques exhaustives observées; par conséquent, l'estimateur du maximum de vraisemblance (EMV) pour θ n'a pas une valeur unique. Pour obtenir une valeur estimée unique pour θ , il faut définir une structure additionnelle.

Une façon d'obtenir un EMV unique est de construire un modèle log-linéaire pour le tableau de contingence élargi en posant comme nulles quelques-unes des interactions de degré supérieur (Little 1985; Fay 1986; Little et Rubin 1987, section 11.6). Nous pourrions tenter de poser comme nulles les interactions qui ne peuvent être estimées à l'aide des données, mais la formalisation qui accompagne cette opération ne se fait pas toujours facilement en pratique. Par exemple, on pourrait croire *a priori* qu'il n'est pas possible d'estimer l'interaction $R_1 n$ est jamais observée lorsque $R_1 \times X_1$ par l'interaction $R_1 \times X_1$ sur l'information de l'intermédiaire d'une

ces distributions pourraient servir à l'estimation ponctuelle ou à l'estimation par intervalle de valeurs présentant un certain intérêt. En pratique, toutefois, ces calculs sont habituellement irréalisables; il faut donc procéder d'une autre façon. Le remplacement des valeurs manquantes par imputation est une solution intéressante car elle produit un ensemble de données complet qui peut être analysé à l'aide de méthodes pour données complètes. Little (1986) expose sommairement les avantages et les inconvénients de diverses méthodes d'imputation; nous n'allons

En pratique, on remplace habituellement chaque valeur manquante par une valeur tirée au hasard dans une distribution; on obtient ainsi un ensemble complet de données que l'on analyse à l'aide des méthodes usuelles pour données complètes. Les estimations d'intervalles établies à l'aide de cette méthode seront artificiellement trop précises puisqu'elles ne reflètent pas la variabilité due à l'imputation. Pour résoudre cette difficulté, on tend de plus en plus à utiliser l'imputation multiple (Rubin 1987), qui consiste à remplacer chaque valeur manquante par m valeurs tirées au hasard dans la distribution. Lorsqu'il y a une quantité modérée de valeurs manquantes, $m = 5$ tirages suffisent pour obtenir des estimations ponctuelles efficaces et des estimations d'intervalles acceptables. Compte tenu des taux de non-réponse observés le plus souvent dans l'EP (en règle générale, 5 à 10 % ou moins, d'après le TOR), $m = 2$ tirages conviendront parfaitement dans la plupart des cas. Toutefois, dans une enquête d'envergure comme l'EP, même un faible nombre de tirages peut poser des problèmes de calcul.

Comme ce qui nous intéresse le plus dans l'EP sont les taux d'appariement estimés pour les strates formées *a posteriori*, il faut vraisemblablement accorder plus d'importance à la variabilité due à l'imputation de X qu'à la variabilité due à l'imputation de Y ; autrement dit, il importe vraisemblablement de faire ressortir la variabilité des taux de sous-dénombrement globaux avant la variabilité des taux de répartition du sous-dénombrement entre les strates formées *a posteriori*. On peut donc obtenir des résultats acceptables en imputant un seul ensemble de valeurs X , puis en imputant Y par multiplication, étant donné X . Une autre solution serait d'imputer un seul ensemble de valeurs X , puis imputer la probabilité d'appariement, étant donné X . C'est ce qui a été fait dans le TOR (Schenker 1988); grâce à cette méthode, les valeurs imputées de X et les valeurs fractionnaires de Y sont considérées comme des valeurs obtenues par imputation simple dans l'estimation des taux de sous-dénombrement. Il se fait actuellement des recherches sur la manière de toujours choisir la méthode d'imputation qui convienne à un ensemble de tableaux de référence donné. Souhaitons que la méthode utilisée dans le TOR, et qui consiste à imputer une seule valeur de X puis à imputer $P(Y = 1 | X)$ s'avère un heureux compromis entre la précision de l'imputation multiple et la facilité d'exécution de l'imputation simple.

3.4 Modèles et méthodes d'estimation

Dans cette section, nous présentons deux méthodes qui servent à modéliser les données manquantes et à estimer les tableaux de référence à des fins d'imputation. Il s'agit de la méthode du maximum de vraisemblance «neutre» (MVN) et de la méthode bayésienne «non neutre» (BN), tout à fait inédite, qui devrait être supérieure à la méthode MVN si la non-réponse n'est pas

3.4.1 Méthode du maximum de vraisemblance «neutre»

Comme nous l'avons indiqué plus tôt, une méthode d'imputation «neutre» exige un seul tableau de référence, qui est appliqué à tous les schémas de non-réponse. Une approche élémentaire est d'estimer ce tableau de référence unique (θ) au moyen des fréquences de cellules observées dans le tableau θ_{CC} . L'estimateur correspondant (θ_{CC} est asymptotiquement sans biais pour θ si la non-réponse est parfaitement indépendante de la valeur de cet élément, que manque un élément de réponse est entièrement indépendante de la valeur de cet élément, que cette valeur soit observée ou manquante. Si la non-réponse est aléatoire, plutôt que parfaitement

indices inférieurs i, j, k , et l désignent les niveaux de X_1, X_2, X_3 , et Y respectivement. Comme nous nous reporterons à Θ' lorsque nous imputerons des valeurs manquantes pour le t -ième tableau de données, nous appellerons Θ' le tableau de référence pour le t -ième tableau de données et $\{\Theta'; t = 1, \dots, 2^4\}$ l'ensemble des tableaux de référence. L'imputation de valeurs manquantes équivaut à élargir chaque tableau supplémentaire de manière à le rendre quadridimensionnel en conformité avec le tableau de référence correspondant. Considérons, par exemple, l'imputation de Y pour les personnes pour lesquelles seule Y est inconnue. Cela revient à élargir le tableau de données pour le schéma de non-réponse 2 en divisant chaque effectif de cellule du tableau en deux nombres, le nombre de personnes pour lesquelles $Y = 1$ et le nombre de personnes pour lesquelles $Y = 0$, les proportions étant conformes à celles du tableau de référence Θ^2 . Comme on connaît Θ^2 , cette opération est simple; nous déterminons tout d'abord à l'aide de Θ^2 la distribution conditionnelle de Y étant donné X pour le schéma de non-réponse en question:

$$(1) \quad P(Y = 1 \mid X_1, X_2, X_3, t = 2) = \frac{\theta_{ijk1}^2}{\theta_{ijk1}^2 + \theta_{ijk0}^2}$$

pour $t = 1, \dots, I, j = 1, \dots, J$, et $k = 1, \dots, K$. Ensuite, nous imputons $Y = 1$ pour chaque observation dans la cellule ijk du tableau de données avec une probabilité égale à la valeur de l'expression définie par le membre de droite de l'équation (1); nous pourrions, à la place, imputer la moyenne de cette distribution, qui correspond précisément à la probabilité d'un appariement (1). Dans la section 3.3, nous examinons les avantages du tirage aléatoire par rapport à ceux de l'imputation de moyenne pour ce qui a trait à l'EP.

Souignons que la seule chose qu'il a fallu déterminer à l'aide de Θ^2 dans l'exemple ci-dessus est la distribution conditionnelle de Y étant donné X ; par conséquent, n'importe quelle valeur de Θ^2 qui produit les mêmes valeurs pour (1) appellera le même processus d'imputation. Ainsi, pour avoir une imputation précise, il n'est pas nécessaire que la valeur estimée de Θ' corresponde à la distribution conjointe de Y et X pour le t -ième schéma de non-réponse; il suffit que la distribution conditionnelle des variables manquantes, étant donné les variables observées, établie à l'aide de la valeur estimée de Θ' se rapproche de la distribution exacte.

En particulier, si la non-réponse est aléatoire, un seul tableau de référence $\Theta' = \Theta$, $t = 1, \dots, 2^4$, produira des valeurs imputées valides pour tous les schémas de non-réponse même si la distribution conjointe de X et Y devait varier d'un schéma à l'autre. Le fait que l'on a besoin d'un seul tableau de référence découle de la définition de la «neutralité», qui implique que la distribution conditionnelle des valeurs manquantes, étant donné les valeurs observées, ne dépend pas du schéma de non-réponse. Les valeurs imputées Θ valides ne proviennent pas de Θ_{CC} , c.-à-d. des probabilités de cellules pour la distribution conjointe de X_1, X_2, X_3 , et Y qui est à la base du tableau CC , mais de la distribution conjointe de X_1, X_2, X_3 et Y , dont on calcule une moyenne pondérée pour l'ensemble des schémas de non-réponse. De façon générale, si la non-réponse est non aléatoire, il faudra définir un tableau de référence propre pour chaque schéma de non-réponse.

Dans notre approche axée sur un modèle, il faudra examiner deux questions essentielles: (1) comment estimer l'ensemble des tableaux de référence en appliquant des principes reconnus de l'estimation efficace et (2) comment exécuter l'imputation une fois que les estimations ont été établies. Dans la section 3.4, nous comparons deux méthodes d'estimation mais avant, nous allons analyser brièvement diverses formes d'imputation.

3.3 Imputation simple ou multiple et imputation de moyenne

Une fois les tableaux de référence estimés, on connaît entièrement les distributions des variables manquantes, étant donné les variables observées, pour chaque personne. En théorie, ces distributions pourraient servir à l'estimation ponctuelle ou à l'estimation par intervalle de

personnes qui se distinguent uniquement par l'origine ethnique, les valeurs des autres variables étant les mêmes pour tout le groupe; il peut être plus difficile d'obtenir des données sur l'origine ethnique dans le cas des groupes minoritaires que dans le cas des groupes non minoritaires; en conséquence, la répartition de la population selon l'origine ethnique variera suivant qu'il s'agit des personnes dont l'origine ethnique est inconnue ou des personnes dont l'origine ethnique est connue. Pareillement, même après avoir tenu compte de toutes les variables X et Z , il y aurait plus de chances que la valeur de X soit manquante pour les personnes qui n'ont pas été recensées que pour celles qui l'ont été. Une méthode d'imputation fondée sur une catégorie générale de modèles «non neutres» est présentée dans la section 3.4.2.

3. AUTRES MÉTHODES D'IMPUTATION POUR L'EP

3.1 Introduction

Soient $X = (X_1, X_2, X_3)$ trois caractéristiques enregistrées par l'EP (par exemple âge, sexe et origine ethnique). Les variables X_1, X_2 , et X_3 sont supposées être qualitatives et peuvent prendre I, J , et K valeurs respectivement. Nous avons choisi trois variables uniquement pour les besoins de l'illustration et pour simplifier la notation; toutes les notions présentées dans cette section s'appliquent automatiquement à n'importe quel nombre de variables qualitatives. Dans la pratique, les variables X comprendront probablement les caractéristiques démographiques et géographiques et les caractéristiques du logement utilisées dans la stratification *a posteriori* pour l'estimation du sous-dénombrement; elles pourront aussi inclure d'autres variables de l'EP, comme le code «déménagement» et le code d'interview, qui ne sont pas d'un intérêt primordial mais qui peuvent être utiles pour l'imputation.

Nous allons former IJK classes de personnes en répartissant ces personnes selon X_1, X_2 , et X_3 . Ces classes peuvent coïncider ou non avec les strates formées *a posteriori* pour l'estimation du sous-dénombrement; en pratique, les strates formées *a posteriori* seront probablement définies plus largement que ces classes. Il est utile mais non nécessaire de définir ces classes sous forme de tableau croisé de toutes les valeurs possibles de X_1, X_2 , et X_3 ; il existe aussi des formes plus complexes (comme les tableaux imbriqués). Nous allons construire des modèles linéaires logarithmiques pour des tableaux à plusieurs entrées ou tableaux de contingence mais ce genre de modèles existe aussi pour d'autres formes de présentation.

Soit Y a variable dichotomique qui désigne le code d'appartenance et qui peut prendre la valeur 1 (appartient résussi) ou la valeur 0 (appartient manqué). S'il ne manquait aucune donnée, on pourrait résumer les résultats de l'EP dans un tableau de contingence quadridimensionnel à $I \times J \times K \times 2$ cellules puisqu'on pourrait classer chaque personne selon X_1, X_2, X_3 , et Y . Or, les personnes pour lesquelles il manque la valeur d'une ou de plusieurs variables ne peuvent être classées que selon les variables qui ont été observées. Les personnes pour lesquelles X_1, X_2, X_3 , et Y ont été observées constitueront un tableau quadridimensionnel que nous appellerons tableau des *cas complets* (CC) ou tableau de données pour le schéma de non-réponse 1 (aucune variable manquante). Les personnes pour lesquelles X_1, X_2, X_3 ont été observées mais non Y constitueront un *tableau supplémentaire* tridimensionnel à IJK cellules, que nous appellerons tableau de données pour le schéma de non-réponse 2. En règle générale, il y aura 2^4 tableaux qui correspondront à tous les schémas de non-réponse possibles, soit un tableau CC et $2^4 - 1$ *tableaux supplémentaires*.

3.2 Imputation à partir de tableaux de référence

D'après notre méthode d'imputation fondée sur un modèle, les tableaux de données seront définis pour différents schémas de non-réponse comme des observations multinomiales. Pour chaque schéma de non-réponse, nous allons définir un ensemble de probabilités de cellules $\Theta^t = \{\Theta^t_{ijk}\}$, où l'indice supérieur t désigne le schéma de non-réponse, $t = 1, \dots, 2^4$, et les

Les valeurs manquantes de X et de Y ont été imputées en deux étapes. (Notre description a été simplifiée dans le but d'alléger la présentation; voir Schenker (1988) pour les détails de la procédure.) Premièrement, on a imputé toutes les valeurs manquantes de X à l'aide d'une méthode «hot deck» fondée sur les variables observées de X ; autrement dit, les valeurs imputées provenaient des distributions observées de X . En deuxième lieu, après avoir imputé les valeurs manquantes de X , on a ajusté un modèle de régression logistique de Y sur X et Z pour les cas où Y avait été observée. Ce modèle de régression logistique a ensuite servi à imputer les probabilités d'appariement pour toutes les valeurs manquantes de Y . On a imputé des probabilités au lieu de zéros et de uns pour a) accroître la précision de l'estimation et b) pouvoir évaluer la variabilité due à l'imputation (Schenker 1989).

2.2 Analyse des méthodes

Les méthodes d'imputation du TOR présentent beaucoup d'avantages. Elles sont faciles à comprendre et utilisent des modèles explicites pour l'imputation de X . En outre, l'imputation se fait surtout en fonction d'observations plutôt qu'en fonction de distributions marginales. Enfin, ces méthodes permettent, en principe, d'évaluer la variabilité des estimations du sous-dénombrement attribuable aux valeurs manquantes de Y . En revanche, ces méthodes présentent certaines lacunes que nous allons décrire ci-dessous.

La méthode d'imputation du TOR est une méthode «neutre» car elle ne tient pas compte du mécanisme de non-réponse. Les méthodes «neutres» supposent que la non-réponse est aléatoire (NRA) (Rubin 1976); en d'autres termes, elles supposent qu'étant donné les observations, le fait qu'il manque des éléments de réponse ne dépend pas de la valeur de ces éléments. Par exemple, si les valeurs de X et de Z sont connues pour toute la population, la NRA suppose qu'on peut imputer la valeur de Y au moyen de la distribution conditionnelle de Y , étant donné X et Z , pour les personnes dont on connaît les valeurs de X , Y et Z .

De fait, la méthode utilisée dans le TOR est un cas particulier d'une méthode «neutre» puisqu'elle renferme des hypothèses qui sont plus rigides que l'hypothèse générale de la non-réponse aléatoire. La méthode du TOR a traité X et Y de façon asymétrique; en effet, selon cette méthode, on a imputé les valeurs manquantes de Y conditionnellement à toutes les valeurs observées alors qu'on a imputé les valeurs de X conditionnellement aux valeurs observées de X plutôt que conditionnellement aux valeurs observées de X , Y et Z . Ainsi, non seulement la méthode du TOR suppose que la non-réponse est aléatoire, mais aussi qu'étant donné les éléments observés de X , les éléments manquants de X sont conditionnellement indépendants de Y et de Z .

Cette autre hypothèse d'indépendance n'est peut-être pas réaliste; il se peut qu'étant donné les valeurs observées de X , il y ait une dépendance résiduelle des valeurs d'éléments manquants de X par rapport à Y ou à Z . Si tel est le cas, les valeurs observées de Y et Z devraient servir à l'imputation des valeurs de X . Supposons, par exemple, que l'on ne connaît pas le sexe d'une personne de l'échantillon de l'EP et que, par ailleurs, les données dont on dispose sur cette personne (par exemple âge, origine ethnique et adresse) ne permettent pas d'appariement avec un enregistrement du recensement ($Y = 0$) et supposons que le taux de sous-dénombrement dans le recensement tend à être plus élevé pour les hommes que pour les femmes qui ont les mêmes caractéristiques. Alors, le fait de savoir que $X = 0$ au lieu de 1 donne à penser qu'il peut s'agir plus d'un homme que d'une femme. La méthode d'imputation «neutre» la plus générale utiliserait l'information fournie par Y et Z pour l'imputation des valeurs manquantes de X ; cela est une des méthodes d'imputation analysées dans la section suivante (voir section 3.4.1).

Un autre aspect de la méthode d'imputation du TOR qui peut être irréaliste est l'hypothèse de «neutralité» proprement dite. La non-réponse n'est peut-être pas aléatoire; autrement dit, étant donné les valeurs observées, le fait qu'il manque des éléments de réponse n'est pas indépendant de la valeur de ces éléments. Si tel était le cas, il conviendrait mieux d'utiliser un modèle «non neutre» pour le mécanisme de non-réponse. Prenons, par exemple, un groupe de

1. Il se peut que l'on n'ait pas toutes les données voulues sur une personne (qu'il s'agisse des caractéristiques géographiques ou démographiques ou encore des caractéristiques du logement); dans ce cas, on ignore à quelle strate formée *a posteriori* appartient cette personne. 2. Après le traitement des données de l'EP, il arrive que l'on n'ait pu déterminer le code d'appariement (variable dichotomique indiquant s'il y a concordance ou non-concordance avec les enregistrements du recensement) ou un code de dénombrement pour certaines personnes. Cela peut se produire, par exemple, lorsque le nom fourni dans l'EP est incomplet ou qu'il est difficile de définir l'adresse où demeurerait la personne le jour du recensement, si cette personne est démenagée depuis.
- Les données manquantes ont été une importante source de variabilité dans l'estimation du taux de sous-dénombrement pour le recensement décennal de 1980 (Freeman et Navidi 1986; Fay, Passell et Robinson 1988, Chapitre 6). Les améliorations apportées au plan de sondage de l'EP devaient avoir pour effet de réduire la quantité de données manquantes en 1990 (Hogan et Wolter 1988); néanmoins, une méthode de traitement des données manquantes s'impose. Pour le Test des opérations de redressement de 1986 (TOR), qui a été créé récemment pour vérifier la validité des méthodes d'estimation du sous-dénombrement et des méthodes de redressement (Diffendal 1988; Schenker 1988), on a utilisé une EP dont le plan de sondage est comparable à celui de l'enquête prévue pour 1990. Dans cet article, nous examinons les méthodes dont on s'est servi dans le TOR (Schenker 1988) pour traiter les données manquantes, définissons les faiblesses probables de ces méthodes et analysons des solutions de rechange. Nous visons essentiellement à signaler les problèmes et à proposer des méthodes pour les résoudre. L'objet de nos recherches à long terme est d'analyser soigneusement ces méthodes. Bien que notre article porte uniquement sur l'imputation de valeurs manquantes dans l'estimation du sous-dénombrement, la non-réponse est aussi le fait de l'échantillon du recensement qui sert à estimer le surdénombrement. Toutefois, comme les problèmes sont les mêmes dans un cas comme dans l'autre (Schenker 1988), notre analyse vaut pour les deux.
- Dans notre analyse des solutions de rechange, nous proposons une méthode inédite fondée sur un modèle de Bayes qui tient compte du mécanisme de non-réponse et qui, par conséquent, ne suppose pas que la non-réponse est aléatoire. Ce genre de modèles pour les données qualitatives incomplètes est un élément nouveau de la théorie du traitement des données manquantes; voir Fay (1986), Little et Rubin (1987, section 11.6) et Baker et Laird (1988) pour des analyses et des comptes rendus d'ouvrages. En outre, le genre de données manquantes dont il est question ici n'est pas propre à l'estimation du sous-dénombrement; le phénomène se retrouve dans beaucoup d'autres situations. Notre analyse porte donc sur le traitement des données qualitatives manquantes dans son ensemble.
- Dans la section 2, nous examinons les méthodes d'imputation utilisées dans le TOR. Dans la section suivante, nous exposons des méthodes de rechange et les illustrons à l'aide d'un exemple simple. Enfin dans la section 4, nous concluons notre analyse.

2. MÉTHODES D'IMPUTATION UTILISÉES DANS LE TOR

2.1 Description des méthodes

Pour chaque personne incluse dans le champ de l'EP, définissons X comme les variables qualitatives pour l'âge, le sexe, l'origine ethnique, le mode d'occupation et le type de construction; définissons Y comme le code d'appariement ($1 =$ concordance, $0 =$ non-concordance) et Z comme les variables qui indiquent si l'on s'agit d'une interview ordinaire ou d'une interview par personne interposée et si la personne en question a déménagé entre le jour du recensement et le jour de l'EP. Dans le TOR, les variables X (sauf le type de construction) ont servi de critères de stratification *a posteriori* (Diffendal 1988); Z a été observée pour tous les membres de l'échantillon de l'EP alors que des valeurs de Y et d'éléments de X étaient parfois manquantes (Schenker 1988).

Méthodes d'imputation de valeurs manquantes dans des enquêtes postcensitaires

DONALD B. RUBIN, JOSEPH L. SCHAFER, et NATHANIEL SCHENKER¹

ABSTRACT

Pour estimer le taux de sous-dénombrement dans le recensement, on exécute une enquête postcensitaire (EP) et on tente d'apparier les enregistrements de cette enquête avec des enregistrements du recensement; le taux d'appariement donne une estimation du taux de couverture du recensement. L'estimation du sous-dénombrement repose sur une stratification *a posteriori* où les caractéristiques géographiques et démographiques et les caractéristiques du logement *X* servent de critères de stratification. Or, la non-réponse fait qu'il manque des données sur *X* pour certaines personnes; en outre, on ne peut déterminer un code d'appariement *Y* pour chaque personne. Il faut donc une méthode pour imputer les valeurs manquantes de *X* et de *Y*. Cet article vise à examiner les méthodes d'imputation qui ont été utilisées dans le Test des opérations de redressement de 1986 (Schenker 1988) et propose deux méthodes de échange axées sur des modèles: (1) une méthode d'estimation de tableau de contingence fondée sur le maximum de vraisemblance, qui ne tient pas compte du mécanisme de non-réponse et (2) une nouvelle méthode d'estimation de tableau de contingence de type bayésien, qui tient compte du mécanisme de non-réponse. La première méthode est plus simple au point de vue du calcul mais la seconde est plus intéressante au point de vue théorique et scientifique.

MOTS CLÉS: Méthodes bayésiennes; données qualitatives; erreur de couverture; algorithme EM; imputation multiple; non-réponse non aléatoire; sous-dénombrement.

1. INTRODUCTION

Depuis un certain temps, le U.S. Bureau of the Census utilise une enquête postcensitaire (EP) pour évaluer l'erreur de couverture dans ses recensements; il prévoit réaliser de nouveau une EP après le recensement décennal de 1990. L'EP vise à déterminer si une personne a été effectivement recensée; pour cela, il faut chercher parmi les enregistrements du recensement celui qui correspondrait à la personne en question (autrement dit, il s'agit d'établir une concordance). La proportion de personnes dans l'EP qui ont été oubliées lors du recensement sert à estimer la proportion de la population qui n'a pas été recensée. Une opération semblable est exécutée lorsqu'on tente d'apparier un échantillon d'enregistrements du recensement avec les enregistrements de l'EP; cela permet d'estimer le surdénombrement dans le recensement, conséquence d'enregistrements erronés (par exemple enregistrements répétés ou fictifs). Les données de l'EP sur les cas de concordance et les enregistrements erronés sont combinées pour estimer la taille de la population au moyen de l'estimateur de système dual; ce genre d'estimateur, fondé sur la méthode de saisie-résaisie, est analysé dans Marks, Selizer et Krotki (1974), Krotki (1978), Wolter (1986), Diffendal (1988) et Fay, Passell et Robinson (1988, Chapitre 5). Les estimations de système dual de la population sont calculées pour des strates définies *a posteriori* selon des caractéristiques géographiques et démographiques (âge, sexe, origine ethnique) et des caractéristiques du logement (mode d'occupation, type de construction). Or, l'EP n'est pas à l'abri de la non-réponse et à ce propos, deux problèmes particuliers viennent compliquer le processus d'estimation:

¹ Donald B. Rubin et Joseph L. Schaffer, Département de statistique, Harvard University, Cambridge, MA 02138, E.-U.; Nathaniel Schenker, Division de la biostatistique, UCLA School of Public Health, Los Angeles, CA 90024, E.-U.

- ERICKSEN, E.P., KADANE, J.B., et TURKEY, J.W. (1987). Adjusting the 1980 census of housing and population. *Technical Report No. 401*, Department of Statistics, Carnegie-Mellon University, Pittsburgh, PA.
- FAY, R.E. III, et HERRIOT, R.A. (1979). Estimates of income for small places: An application of James-Stein to census data. *Journal of the American Statistical Association*, 74, 269-277.
- FERGUSON, T.S. (1967). *Mathematical Statistics: A Decision Theoretic Approach*. New York: Academic Press.
- FREEDMAN, D.A., et NAVIDI, W.C. (1986). Regression models for adjusting the 1980 census. *Statistical Science*, 1, 3-39.
- GOLDSTEIN, M. (1975). Approximate Bayes solutions to some nonparametric problems. *Annals of Statistics*, 3, 512-517.
- HENDERSON, C.R. (1976). A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics*, 32, 69-83.
- HUI, S.L., et BERGER, J.O. (1983). Empirical Bayes estimation of rates in longitudinal studies. *Journal of the American Statistical Association*, 78, 753-760.
- ISAKI, C.T., DIFFENDAL, G.J., et SCHULTZ, L.K. (1986). Statistical synthetic estimates of undercount for small areas. *Proceedings of Bureau of the Census Second Annual Research Conference*. Bureau of the Census, Washington, D.C., 557-569.
- KADANE, J.B. (1984). Allocating Congressional seats among the states when state populations are uncertain. *Technical Report No. 309*, Department of Statistics, Carnegie-Mellon University, Pittsburgh, PA.
- LINDLEY, D.V., et SMITH, A.F.M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Sér. B*, 34, 1-41.
- LOUIS, T.A. (1984). Estimating a population of parameter values using Bayes and empirical Bayes methods. *Journal of the American Statistical Association*, 79, 393-398.
- MORRIS, C.N. (1983). Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association*, 78, 47-55.
- MULRY-LIGGAN, M., et HOGAN, H. (1986). Research plan on census adjustment standards. *Proceedings of Bureau of the Census Second Annual Research Conference*. Bureau of the Census, Washington, D.C., 381-392.
- NATIONAL ACADEMY OF SCIENCES (1985). *The Bicentennial Census: New Directions for Methodology in 1990*, (eds. C.F. Citro et M.L. Cohen). Washington: National Academy Press.
- READ, T.R.C., et CRESSIE, N.A.C. (1988). *Goodness-of-fit Statistics for Discrete Multivariate Data*. New York: Springer-Verlag.
- SCHULTZ, L.K., HUANG, E.T., DIFFENDAL, G.J., et ISAKI, C.T. (1986). Some effects of statistical synthetic estimation on census undercount of small areas. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 321-325.
- STROUD, T.W.F. (1987). Bayes and empirical Bayes approaches to small area estimation, dans *Small Area Statistics*, (eds. R. Plalek, J.N.K. Rao, C.E. Särndal et M.P. Singh). New York: Wiley, 124-137.
- TURKEY, J.W. (1981). Discussion sur, "Issues in adjusting for the 1980 census undercount," par Barbara Bailar et Nathan Keyfitz, à la réunion annuelle de l'American Statistical Association, Detroit, MI.

REMERCIEMENTS

pourrait modifier l'ordre des termes de l'inéquation (3.28). Pour ce qui est des Etats toute-fois, Cressie (1988) et le tableau 1 montrent par une analyse des résiduels que (2.7) et (2.10) conviennent aux données PEP 3-8 de 1980, et comme (3.29) implique que la condition (4.1) est satisfaisante, on peut penser avec suffisamment de certitude que l'estimateur empirique de Bayes ordinaire produira, pour chaque Etat, des chiffres de population plus justes que le recensement. Cela pourrait ne pas être le cas au niveau de l'ilot; il est clair qu'il faudra décider du niveau où l'on juge primordial d'avoir des chiffres de population exacts. L'objectif premier du recensement aux E.-U. est de fournir les chiffres de population des Etats au Congrès en vue de la redistribution des sièges à la Chambre des représentants. On pourrait, à cet égard, inclure un certain nombre de grandes villes parmi les Etats et considérer, par exemple, la ville de New York comme un "Etat" et l'Etat de New York moins la ville de New York comme un autre Etat. Il nous semble que ce niveau d'aggrégation est le plus névralgique politiquement et que les recherches devraient être axées en priorité sur l'établissement de chiffres exacts pour ce niveau.

Pour cette étude, l'auteur a bénéficié de la collaboration de membres du Undercount Research Staff du U.S. Bureau of the Census et des commentateurs perspicaces de J.B. Kadane et de J.W. Tukey. Cette étude a aussi été rendue possible grâce aux conventions n° 86-5, 87-4, 87-10 et 88-13 sur la statistique entre l'Université Iowa State et le Census Bureau (Joint Statistical Agreements). Les opinions exprimées dans cet article sont celles de l'auteur et ne reflètent pas nécessairement la position du Census Bureau.

BIBLIOGRAPHIE

COWAN, C.D., et BETTIN, P.J. (1982). Estimates and missing data problems in the post enumeration survey. Rapport interne, Statistical Methods Division, Bureau of the Census, Washington, D.C.

CRESSIE, N. (1986). Empirical Bayes estimation of undercount in the decennial census. *Statistical Laboratory Preprint 86-58*, Iowa State University, Ames, IA.

CRESSIE, N. (1987a). Empirical Bayes estimation of undercount in the decennial census. Document soumis au *Journal of the American Statistical Association*.

CRESSIE, N. (1987b). Commentaire sur, "Census undercount adjustment and the quality of geographic population distributions," par A.L. Schirm et S.H. Preston. *Journal of the American Statistical Association*, 82, 980-983.

CRESSIE, N. (1988). Estimating census undercount at national and subnational levels. *Proceedings of Bureau of the Census Fourth Annual Research Conference*. Bureau of the Census, Washington, D.C., 123-150.

CRESSIE, N., et DAJANI, A. (1988). Empirical Bayes estimation of U.S. undercount based on artificial populations. *Statistical Laboratory Preprint 88-17*, Iowa State University, Ames, IA.

DEMPTSTER, A.P., et TOMBERLIN, T.J. (1980). The analysis of census undercount from a post-enumeration survey, dans *Proceedings of the 1980 Conference on Census Undercount*. Bureau of the Census, Washington, D.C. 88-94.

ERICKSEN, E.P., et KADANE, J.B. (1985). Estimating the population in a census year: 1980 and beyond. *Journal of the American Statistical Association*, 80, 98-131.

ERICKSEN, E.P., et KADANE, J.B. (1987). Sensitivity analysis of local estimates of undercount in the 1980 U.S. Census, dans *Small Area Statistics*, (éds. R. Platek, J.N.K. Rao, C.E. Särndal et M.P. Singh.) New York: Wiley, 23-45.

qui est similaire à (2.15). Il est clair que la suite d'inéquations (3.28) se vérifie dans les mêmes conditions qu'apparavant, soit (3.22), (3.24) et (3.26); nous considérons que est-risk_j dans (3.28) est égal à (3.30) où $r_j = D_j$ lorsque "est" = "ueb," $r_j = D_j$ lorsque "est" = "ceb," et $r_j = 0$ lorsque "est" = "syn". De plus, pour ce qui a trait à la fonction de perte (2.15), où $f(C_j) = 1/C_j$, la différence de risque augmente à mesure que l'on descend dans la hiérarchie géographique.

4. ANALYSE

La triple inéquation (3.28) repose sur diverses hypothèses qu'il convient d'approfondir. On a supposé que les modèles (2.7) et (2.10) convenaient et, en particulier, que les distributions étaient indépendantes d'une sous-région à l'autre. Nous avons aussi supposé que l'effet de l'estimation de D_j dans les estimateurs empiriques de Bayes de F_{ji} était négligeable. Signalons toutefois que les estimateurs synthétiques de F_{ji} n'utilisent pas d'estimation D_j et que par conséquent, la suite d'inéquations (3.28) dépend uniquement de la validité des modèles (2.7) et (2.10).

Les conditions qui sous-tendent la triple inéquation (3.28) se réaliseront dans la mesure où σ_j^2/τ_j^2 sera "faible". Cela implique qu'il faudra choisir un grand nombre de ménages dans l'enquête post-censitaire (EP) sinon on ne pourra être assuré de l'efficacité des opérations de redressement. En connaissant la variance de strate (d'après les résultats d'un recensement antérieur par exemple), on pourrait *élaborer* le plan de l'EP de manière que les conditions soient satisfaites. Une fois l'enquête réalisée et les données $\{X_{ji}: i = 1, \dots, I; j = 1, \dots, J\}$ connues, on peut vérifier les conditions (3.22), (3.24) et (3.26) au moyen des estimateurs $\hat{\tau}_j^2$ et $\hat{\sigma}_j^2$ définis en (3.2).

Concentrons notre attention sur la meilleure combinaison convexe de X_{ji} et de $X_{j\cdot}$, soit $F_{j\cdot}^{\text{ueb}}$ défini en (3.3). Alors, $\text{ueb-risk}_j \leq \text{cen-risk}_j$, si (3.22) se vérifie, c'est-à-dire si

$$\left\{ \sum^h C_{jh}/C_j \right\}^{1/2}; j = 1, \dots, J. \quad (4.1)$$

Il convient de souligner que cette condition est moins stricte lorsque la région i compte une faible population recensée; par contre, pour ce qui est des régions qui ont une population recensée élevée, l'estimateur empirique de Bayes ordinaire peut produire une estimation de la population moins juste que les chiffres du recensement. Une condition suffisante pour que (4.1) se vérifie est $\sigma_j^2/\tau_j^2 \leq 1; j = 1, \dots, J$, cette condition est également celle qui est requise pour que l'estimateur synthétique produise une estimation de la population plus juste que les chiffres du recensement. Elle a, de fait, été satisfaite pour les données PEP 3-8 de 1980 (voir Section 3.2).

Enfin, la condition (4.1) devient moins stricte aux niveaux inférieurs et, de fait, les résultats de la Section 3.2 indiquent que la différence de risque entre les chiffres redressés et les chiffres bruts s'accroît. Examinons cela de plus près. Les résultats sont justes à la condition que le modèle vaille pour les niveaux inférieurs, mais cela n'est probablement pas le cas au niveau de l'ilot et du secteur de dénombrement. L'existence d'un biais dans (2.7) et (2.10) à savoir

$$E(F_{ji}) = F_j + b_{ji}; E(X_{ji} | F_{ji}) = F_{ji} + d_{ji}, \quad (4.2)$$

alors

(3.25) $\text{syn-risk}_i \leq \text{cen-risk}_i.$

Enfin, si $(\sigma_j^2/\tau_j^2) \leq 1$, et

(3.26)
$$4(\sigma_j^2/\tau_j^2)^2 \left(\frac{\sum_{jh}^h C_{jh}}{C_{ji}} \right)^2 - (\sigma_j^2/\tau_j^2) \left(1 + \frac{\sum_{jh}^h C_{jh}}{2C_{ji}} \right) + 3 \geq 0; j = 1, \dots, J,$$

alors

(3.27) $\text{ceb-risk}_i \leq \text{syn-risk}_i.$

encore une fois (selon (3.26)), si σ_j^2/τ_j^2 est faible, les risques peuvent être bornés.

Par conséquent, il est possible d'établir une inéquation analogue à (2.32):

(3.28) $\text{ueb-risk}_i \leq \text{ceb-risk}_i \leq \text{syn-risk}_i \leq \text{cen-risk}_i,$

où l'inégalité du centre exige la réalisation de la condition (3.26) et l'inégalité de droite exige la réalisation de la condition (3.24). Si l'une ou l'autre de ces inégalités ne se vérifie pas, nous pouvons au moins affirmer que l'estimateur empirique de Bayes ordinaire représente une amélioration par rapport aux chiffres du recensement si la condition (3.22) est réalisée. Pour les données PEP 3-8 du recensement de 1980, nous avons

(3.29) $\hat{\sigma}_1^2/\hat{\tau}_1^2 = 0.77, \hat{\sigma}_2^2/\hat{\tau}_2^2 = 0.80, \hat{\sigma}_3^2/\hat{\tau}_3^2 = 1.00;$

autrement dit, en ce qui concerne le recensement décennal de 1980 aux E.-U., le risque lié à l'utilisation des chiffres du recensement est plus élevé que celui lié à l'utilisation de l'estimateur synthétique et le risque lié à l'utilisation de l'estimateur empirique de Bayes ordinaire est le moins élevé de tous.

Comparons maintenant le risque lié à l'utilisation de $Y_{\text{sys}}^{i_1}$ et de $Y_{\text{sys}}^{i_2}$ (estimateurs respectifs de Y_{i_1} et de Y_{i_2} fondés sur F_{est}^{ji} défini en (3.17)) au risque lié à l'utilisation de C_{i_1} et de C_{i_2} , où la région $i = i_1$ & i_2 est composée de deux régions distinctes i_1 et i_2 .

$$\begin{aligned} & \left[\sum_{j=1}^J E \left(Y_{i_1}^{j_2} - Y_{i_2}^{j_2} f(C_{i_1}) \right) \right] \\ &= \sum_{j=1}^J \sum_{f=1}^f \left[\tau_j^2 \left(1 - r_j \right) \frac{1}{C_{ji}} - \frac{\sum_{jh}^h C_{jh}}{1} + \left(\frac{C_{ji}}{1} - \frac{1}{C_{ji}} \right) \right] \\ &+ \left\{ \frac{1}{1 - r_j^2} \sum_{f=1}^f C_{jh}^h + \frac{C_{ji}^2}{r_j^2} \right\} \left[C_{ji}^2 f(C_{i_1}) \right]. \end{aligned}$$

(3.30)

Nous nous servirons de l'équation (3.16) pour comparer les estimateurs identifiés par "est" = "neb," "ceb," et "syn," à "cen" (F_{cen}^{jj}) via (3.16). Considérons l'estimateur

$$F_{est}^{jj} = r_j X_{jj} + (1 - r_j) X_{j\cdot}; 0 \leq r_j \leq 1, \quad (3.17)$$

une combinaison convexe des données X_{jj} et de l'estimateur synthétique $X_{j\cdot}$. Alors,

$$est-risk_j = \sum_{j=1}^J \tau_j^2 (1 - r_j) {}_2C_{jh} - \frac{{}_2C_{jh}^2}{C_{jh}^2} \left\{ r_j^2 C_{jh} + a_j^2 \right\} + \frac{\sum_{j=1}^h C_{jh}}{(1 - r_j^2) C_{jh}^2}. \quad (3.18)$$

Il est facile de voir que la valeur de r_j qui minimise (3.18) est $r_j = D_j = \tau_j^2 / (\tau_j^2 + a_j^2)$; autrement dit, si nous ne tenons pas compte de l'effet de l'estimation de τ_j^2 et de a_j^2 , nous obtenons

$$neb-risk_j \leq est-risk_j; 0 \leq r_j \leq 1. \quad (3.19)$$

Comparons maintenant $neb-risk_j$ (posons $r_j = D_j$ dans (3.17)) et $cen-risk_j$; or selon (2.29)

$$cen-risk_j = \sum_{j=1}^J \tau_j^2 {}_2C_{jh} f(C_j) + \left[\sum_{j=1}^J (F_j - 1) C_{jh} \right] {}_2f(C_j). \quad (3.20)$$

De plus, en posant $\tau_j^2 = k_j a_j^2$; $j = 1, \dots, J$,

$$neb-risk_j = \sum_j a_j^2 \left\{ \frac{1 + k_j}{k_j} + \frac{\sum_{j=1}^h C_{jh}}{C_{jh}} \cdot \frac{1 + k_j}{1} \right\} C_{jh} f(C_j). \quad (3.21)$$

Une condition suffisante pour que $neb-risk_j \leq cen-risk_j$ est,

$$\left\{ \frac{1 + k_j}{k_j} + \frac{\sum_{j=1}^h C_{jh}}{C_{jh}} \cdot \frac{1 + k_j}{1} \right\} \leq k_j;$$

autrement dit, si

$$a_j^2 / \tau_j^2 \leq \sum_{j=1}^h C_{jh} / C_{jh} \left\{ \frac{1 + k_j}{k_j}; j = 1, \dots, J, \right. \quad (3.22)$$

alors

$$neb-risk_j \leq cen-risk_j. \quad (3.23)$$

De même, on peut montrer que si

$$a_j^2 / \tau_j^2 \leq 1; j = 1, \dots, J, \quad (3.24)$$

Se servant des données PEP 3-8 et des équations (3.1) et (3.2), Cressie (1987a) a estimé la moyenne de la distribution composée de même que les variances de strate et d'échantillonnage uniformisées définies dans (2.7) et (2.10):

$$(3.11) \quad \text{noirs:} \quad F_1 = 1.06076 \quad \hat{r}_1^2 = 673.982 \quad \hat{\sigma}_1^2 = 522.183,$$

$$(3.12) \quad \text{non-noirs hispaniques:} \quad F_2 = 1.04667 \quad \hat{r}_2^2 = 308.990 \quad \hat{\sigma}_2^2 = 246.585,$$

$$(3.13) \quad \text{autres:} \quad F_3 = 0.99981 \quad \hat{r}_3^2 = 242.134 \quad \hat{\sigma}_3^2 = 242.152.$$

Se fondant sur ces estimateurs de paramètres et les données PEP 3-8 $\{X_{ji}: j = 1, 2, 3; i = 1, \dots, 51\}$, Cressie (1987a, tableaux 2 et 3) a produit des estimations du sous-dénombrément $\{u_{jst}^{ji}\}$, $\{u_{jst}^{jst}\}$ pour les estimateurs empiriques de Bayes ordinaires et les estimateurs synthétiques définis par (3.3) et (3.7) respectivement.

Pour vérifier l'ajustement du modèle, on a calculé les résiduels $\{C_{ji}^{ji}(F_{jst}^{jst} - F_{jst}^{jst})\}: i = 1, \dots, I\}$ pour chacune des strates. Les résultats pertinents figurent en affichage arborescent dans le tableau 1; un diagramme en cloche est ce qui illustre le mieux la tendance des résultats pour chaque strate. Le modèle semble être conforme aux données, sauf en ce qui concerne la strate des non-noirs hispaniques dans l'Etat de New York. Eu égard à la cause Cuomo c. Baldrige, la plaidée devant la Southern District Court de New York en 1983, ce nouveau mode de présentation des données révèle des détails intéressants. Les non-noirs hispaniques de l'Etat de New York étaient fortement sous-dénombrés, même par rapport aux non-noirs hispaniques des autres Etats. De fait, le juge a tranché en faveur du Département du commerce des E.-U. (en décembre 1987) en faisant valoir que les statisticiens et les démographes n'avaient pas encore mis au point, en 1980, des méthodes de redressement qui pouvaient s'appliquer convenablement à l'échelle du pays.

Dans quelles circonstances les chiffres du recensement sont-ils améliorés lorsqu'on remplace $\{C_{ji}: i = 1, \dots, I\}$ par $\{Y_{jst}^{jst}: i = 1, \dots, I\}$? Dans la section suivante, nous définissons les conditions dans lesquelles l'inégalité (2.32) se vérifie dans un modèle empirique de Bayes.

3.2 Redressement à divers niveaux; paramètres de modèle estimés

Les remarques qui ont été faites au début de la Section 2.3 s'appliquent également ici; dans une approche fondée sur un modèle, l'existence d'un faible risque ne garantit pas une faible perte dans tous les cas mais seulement en moyenne. En outre, la propriété d'agrégation définie en (2.24) vaut aussi pour les estimateurs empiriques de Bayes ordinaires (ueb) et conditionnels (ceb) et les estimateurs synthétiques (syn);

$$(3.14) \quad Y_{jst}^{jst} + Y_{jst}^{jst} = Y_{jst}^{jst}$$

car "est" = "ueb," "ceb," et "syn," définis respectivement par (3.4), (3.6) et (3.8). De plus, la propriété d'agrégation-désagrégation définie en (2.27), à savoir

$$(3.15) \quad Y_{jst}^{jst} + Y_{jst}^{jst} = Y_{jst}^{jst},$$

où $i = i_1 \& i_2$ et $F_{jst}^{jst} = F_{jst}^{jst} = F_{jst}^{jst}$, vaut pour n'importe quel estimateur de F_{ji} , y compris les estimateurs empiriques de Bayes ordinaires et conditionnels et les estimateurs synthétiques. Définissons le risque lié à l'estimation de Y_i par $Y_{jst}^{jst} (= \sum_{j=1}^J F_{jst}^{jst} C_{ji}^{jst})$ comme

$$(3.16) \quad \text{est-risk}_i \equiv E[(Y_{jst}^{jst} - Y_i)f(C_i)].$$

Tableau 1
Diagramme arborescent des résiduels fondés sur l'estimateur empirique de Bayes conditionnel (ceb)

Noirs (<i>j</i> = 1)				Non-noirs hispaniques (<i>j</i> = 2)				Autres (<i>j</i> = 1)			

$$F_{\text{ceb}}^{ji} = X_{ji} + \{ \tau_j^2 / (\tau_j^2 + \sigma_j^2) \} \frac{1}{2} (X_{ji} - X_{ji}), \quad (3.5)$$

$$Y_{\text{ceb}}^i = \sum_{j=1}^J F_{\text{ceb}}^{ji} C_{ji}; i = 1, \dots, I. \quad (3.6)$$

L'estimateur empirique de Bayes ordinaire (3.3) peut également se déduire de la théorie des modèles linéaires avec effets aléatoires (Henderson 1976).

Il convient de souligner que les estimateurs empiriques de Bayes des facteurs de redressement de la strate j se ramènent tous à l'estimateur synthétique X_{ji} . Lorsque $\tau_j^2 = 0$. La présence du coefficient de pondération $\{ \tau_j^2 / (\tau_j^2 + \sigma_j^2) \} \frac{1}{2}$ dans la formule de l'estimateur empirique de Bayes conditionnel (3.5) peut paraître étrange à première vue mais Cressie (1987a) montre que ce coefficient donne un estimateur sans biais de l'erreur de strate $C_{ji}^{\text{ce}} (F_{ji} - F_j)$.

Dempster et Tomberlin (1980) avaient proposé quelques années auparavant une autre façon d'estimer le sous-dénombrement à l'aide d'un modèle empirique de Bayes; ils ont en effet avancé que le nombre de personnes non recensées dans une sous-région pourrait être une variable aléatoire binomiale. Ils ont défini un modèle de Bayes hiérarchique mais n'ont pas tenu compte de la variation hétéroscédastique. Stroud (1987) introduit une covariable dans un modèle bayésien à deux degrés mais ses hypothèses (à savoir, variation homoscedastique et échantillons de même taille dans toutes les sous-régions) sont trop restrictives pour le problème qui nous intéresse ici. Cressie (1987a, Section 4) donne les formules du biais et de l'erreur quadratique moyenne pour les estimateurs empiriques de Bayes ordinaires (ueb) (équ. 3.3 et 3.4) et conditionnels (ceb) (équ. 3.5 et 3.6) et les estimateurs synthétiques

$$F_{\text{syn}}^{ji} = X_{ji}. \quad (3.7)$$

$$Y_{\text{syn}}^i = \sum_{j=1}^J F_{\text{syn}}^{ji} C_{ji}; i = 1, \dots, I, \quad (3.8)$$

Comme le taux de sous-dénombrement est une fonction non linéaire de la population réelle, ses estimateurs fondés sur $\{ F_{\text{est}}^{ji}; i = 1, \dots, I; j = 1, \dots, J \}$, notamment

$$u_{\text{est}}^{ji} \equiv 1 - \frac{1}{F_{\text{est}}^{ji}}; i = 1, \dots, I; j = 1, \dots, J, \quad (3.9)$$

$$u_{\text{est}}^i \equiv 1 - \frac{C_i}{Y_{\text{est}}^i}; i = 1, \dots, I, \quad (3.10)$$

sont biaisés; on peut estimer les biais et les erreurs quadratiques moyennes par la méthode de Cressie 1987a, Section 4). Tous ces calculs ne tiennent pas compte de la variabilité due à l'estimation (non linéaire) de $\tau_j^2 / (\tau_j^2 + \sigma_j^2)$.

Supposons que l'on choisit les trois strates suivantes (formées selon l'origine raciale ou ethnique des membres de la population des États-Unis): noirs, non-noirs hispaniques et autres. Les données de l'enquête post-censitaire qui a suivi le recensement de 1980 sont reproduites dans Cressie (1987a, tableau 1). Ces données ont trait à la population hors établissement (*institutionnel*) (Cowan et Betrin 1982) et ont été identifiées "PEP 3-8" par le U.S. Census Bureau - le chiffre 3 désigne le groupe de personnes non recensées dont on a pu estimer le nombre grâce à une enquête réalisée en avril et à l'imputation de données, tandis que le chiffre 8 désigne les enregistrements erronés qui ont pu être portés à l'attention des autorités grâce à une enquête indépendante qui a permis l'imputation de données à partir de renseignements fournis par le U.S. Post Office.

3. REDRESSEMENT DES CHIFFRES DU RECENSEMENT PAR L'ESTIMATEUR EMPIRIQUE DE BAYES

Nous avons défini plus tôt par les équations (2.14), (2.21) et (2.5) les estimateurs de la population réelle Y'_{uba} , Y'_{dba} , et Y'_{sya} , pour la région i . Pour que ces estimateurs soient fonction uniquement des données, il nous faut des estimateurs pour les paramètres inconnus F_j , τ_j^2 , et σ_j^2 ; Fay et Herriot (1979) définissent des estimateurs empiriques de Bayes dans un modèle de régression dont les modèles (2.7) et (2.10) sont des cas particuliers. Pour des raisons de cohérence statistique (voir Cressie 1987, Section 3.3), nous choisissons

$$F_j = X_j. \quad (3.1)$$

$$\tau_j^2 = \max \left\{ \left[\sum_{i=1}^I C_{ji} I(C_{ji} > 0) (X_{ji} - X_j)^2 / \left(\sum_{i=1}^I I(C_{ji} > 0) - 1 \right) \right] - \hat{\sigma}_j^2, 0 \right\}. \quad (3.2)$$

La valeur de $\hat{\sigma}_j^2$ est obtenue à la suite d'un échantillonnage: on la calcule par l'estimation de système dual et Schultz et coll. (1986) la déterminent pour leurs populations fictives en répétant l'échantillonnage de 1,440 secteurs de dénombrement (parmi les quelque 300,000 qui existent) avec probabilité proportionnelle à la taille.

La stabilité statistique (c'est-à-dire, faible variance d'échantillonnage) est plus facile à obtenir pour les moyennes d'échantillon que pour les variances d'échantillon. Le coefficient de variation de la variance de l'échantillon est environ $\sqrt{2/\sqrt{n}}$; par conséquent, pour obtenir un intervalle de confiance relatif (0.5, 1.5) pour la variance de la population, nous devons avoir $n = 32$ et pour obtenir un intervalle (0.95, 1.05), il nous faut $n = 3,200$. L'estimateur $\sum_{i=1}^I C_{ji} I(C_{ji} > 0) (X_{ji} - X_j)^2 / (\sum_{i=1}^I I(C_{ji} > 0) - 1)$ de $\tau_j^2 + \sigma_j^2$ est donc très instable, surtout lorsqu'il y a beaucoup de strates et que, par conséquent, $\sum_{i=1}^I I(C_{ji} > 0)$ est faible (inférieur à 30).

Une façon de contourner la difficulté serait de définir une autre distribution composée, notamment en stipulant que l'ensemble $\{\tau_j^2: j = 1, \dots, J\}$ est issu, par exemple, de la répartition gamma. Ainsi, au lieu d'estimer J paramètres $\{\tau_j^2: j = 1, \dots, J\}$, nous n'aurions qu'à estimer deux paramètres gamma (voir par exemple Hui et Berger 1983). Une autre solution possible serait de grouper provisoirement quelques-unes des strates dans le but d'estimer la variance de strate. Autrement dit, il s'agirait de définir des groupes d'indices de strates distincts, A_1, \dots, A_K , de telle sorte que $\cup \{A_k: k = 1, \dots, K\} = \{1, 2, \dots, J\}$, et $\tau_j^2 = \tau_{j'}^2 = T_k^2$, lorsque j et j' se rapportent au même A_k . De cette manière, Cressie et Dajani (1988) réussissent à faire passer le nombre de paramètres de la variance de strate de $J = 96$ à $K = 4$. Pour ce qui a trait aux données analysées ci-dessous, il n'a pas été nécessaire de procéder de cette façon puisque $\sum_{i=1}^I I(C_{ji} > 0) = 51$ pour chacune des trois strates formées selon l'origine raciale.

3.1 Estimateurs empiriques de Bayes

Nous pouvons maintenant construire les estimateurs empiriques de Bayes ordinaires (voir p. ex., Morris 1983) et conditionnels (Louis 1984):

$$F_j^{\text{néb}} = X_j + \{\tau_j^2 / (\tau_j^2 + \hat{\sigma}_j^2)\} (X_{ji} - X_j), \quad (3.3)$$

$$Y_j^{\text{néb}} = \sum_j F_j^{\text{néb}} C_{ji}; \quad i = 1, \dots, I; \quad (3.4)$$

En l'absence d'autres considérations (p. ex. considérations politiques, pratiques, etc.), il est juste et normal d'utiliser la méthode qui présente le moins de risques. Le statisticien sait que ce *mode d'opération* produira en moyenne de meilleures estimations, si l'on considère ici l'ensemble des problèmes auxquels s'intéresse le statisticien. Cependant, rien ne nous assure que, pour un problème donné (en l'occurrence l'estimation du taux de sous-dénombrement dans le recensement de 1990), la perte sera moins élevée pour une série d'estimations fondées sur le critère du risque minimum que pour une autre série d'estimations. En d'autres termes, l'inéquation $E(V^2) < E(W^2)$ ne garantit pas que $V^2 < W^2$ pour un problème particulier. Si, à la lumière des données recueillies, nous devons reconnaître que la méthode choisie n'est pas celle qui présente le moins de risques, nous devrions néanmoins la considérer comme optimale. Dans le reste de cette section, nous présentons divers résultats concernant les estimateurs de Bayes (pour les démonstrations, voir Cressie 1988). Il va sans dire que ces résultats dépendent de la justesse du modèle proposé. En pratique, les résultats les plus pertinents ont trait aux estimateurs *empiriques* de Bayes; ces résultats sont reproduits dans la Section 3 (avec démonstrations).

Ce qu'il faut surtout se rappeler à propos des estimateurs de Bayes ordinaires (2.13) et (2.14) (voir Section 2.2) est qu'ils sont optimaux ou quasi-optimaux pour une grande catégorie de fonctions de perte. En outre, ils sont indépendants des niveaux, c'est-à-dire qu'ils sont optimaux non seulement au niveau où ils sont construits, mais aussi aux niveaux supérieurs, lorsqu'il y a agrégation. Selon (2.14),

$$Y_{nba}^i + Y_{nba}^{i&I'} = Y_{nba}^{i&I'}, \tag{2.24}$$

où $i&I'$ désigne la région formée des deux régions distinctes i et I' . Ainsi, on devrait tenter de construire un estimateur de Bayes au niveau d'agrégation le plus bas (lots de recensement), puis remonter jusqu'au niveau voulu, ce qui assurerait la cohérence des données à tous les niveaux. Dans la pratique, une telle chose est impensable car l'échantillon de l'enquête post-censitaire ne sera *jamais* assez grand pour produire des estimations de système dual du taux de sous-dénombrement pour tous les lots. Il en est de même pour les secteurs de dénombrement et les comtés. De plus, à ces niveaux les modèles (2.7) et (2.10) ne conviennent plus aussi bien (Cressie et Dajani 1988); dans la Section 3.1, il est question d'un ajustement convenable au niveau de l'Etat.

Il est certain que l'enquête post-censitaire produira des renseignements sur chacun des 51 Etats, ce qui permettra la construction d'estimateurs (empiriques) de Bayes au niveau de l'Etat. Ce niveau est le plus névralgique politiquement; les chiffres des recensements décennaux servent en premier lieu à la redistribution des sièges des 50 Etats (hormis Washington, D.C.) à la Chambre des représentants (ces chiffres doivent parvenir au Congrès au plus tard le 31 décembre de l'année du recensement). À ce niveau, les estimateurs de Bayes (2.13) et (2.14) offrent donc un compromis entre les facteurs de redressement observés d'un Etat $\{X_{ji}^j: j = 1, \dots, J\}$; et les facteurs de redressement (synthétiques) $\{F_{ji}^j: j = 1, \dots, J\}$. Par exemple, lorsqu'on utilise les estimateurs de Bayes, on reconnaît que le taux de sous-dénombrement chez les noirs du Mississippi peut être différent de celui observé chez les noirs de New York.

Nous allons maintenant examiner les conséquences de l'estimation *synthétique* aux niveaux inférieurs après qu'une estimation bayésienne a été faite à un niveau donné. Pour assurer la cohérence des données à tous les niveaux, il est souhaitable d'estimer le taux de sous-dénombrement au niveau de l'lot, puis de remonter jusqu'au niveau voulu. Supposons que nous estimions un facteur de redressement $F_{ji}^{I'}$ pour la strate j dans la région I' . Maintenant, supposons que $I = I_1 \& I_2$, c'est-à-dire que la région I est composée de deux sous-régions distinctes I_1 et I_2 . Alors la méthode synthétique, aux niveaux inférieurs, pose que

$$F_{ji}^{I'} = F_{ji}^{I_2} = F_{ji}^{I_1}, \tag{2.25}$$

de sorte que les estimateurs de la population réelle sont définis,

Nous venons de démontrer que les estimateurs (2.13) et (2.14) sont des estimateurs de Bayes (ou du type Bayes) pour une grande catégorie de fonctions de perte. Cependant, il n'est pas sûr que les propriétés d'ensemble de $\{F_{jba}^{ji}; i = 1, \dots, I; j = 1, \dots, J\}$, sont de parfaits estimateurs des propriétés d'ensemble correspondantes de $\{F_{ji}^{ji}; i = 1, \dots, I; j = 1, \dots, J\}$, à cause de l'inégalité $\text{var}(\theta) \geq \text{var}(E(\theta | X))$. Selon cette inéquation, la variance de la moyenne a posteriori du paramètre est moindre que la variance du paramètre proprement dit. Cela ne change rien à l'estimation de la population des Etats mais pour ce qui est d'estimer la *distribution* de $\{F_{ji}^{ji}; i = 1, \dots, I; j = 1, \dots, J\}$, ou $\{X_j; i = 1, \dots, I\}$, (2.13) par exemple, l'estimateur (2.13) est peu approprié. Ce genre de distribution est utilisée dans les études de normes (Mulry-Liggan et Hogan 1986) pour déterminer la proportion de personnes dans une strate qui sont touchées par un taux de sous-dénombrement supérieur à u% (Cressie 1988, Section 4).

Nous allons imposer une contrainte à l'estimateur de $\{F_{ji}^{ji}; i = 1, \dots, I\}$ de manière que les moments a posteriori des fonctions de distribution empiriques (pondérées) de cet estimateur correspondent aux moments de sa fonction de distribution empirique pondérée. Pour cela, il suffit de modifier l'estimateur de Bayes ordinaire pour obtenir un estimateur de Bayes conditionnel ayant les propriétés d'ensemble appropriées. Louis (1984) expose en détail une version des modèles (2.7) et (2.10) à variances égales; néanmoins, cette approche peut être facilement modifiée pour des variances pondérées. Cressie (1986) montre que l'estimateur de

$$F_{jba}^{ji} = \zeta_j + G_j(X_{ji} - \zeta_j), \quad (2.20)$$

$$Y_{jba}^{ji} = \sum_j F_{jba}^{ji} C_{ji}, \quad (2.21)$$

la première équation étant obtenue en résolvant les équations ci-dessous en fonction de ζ_j et de G_j :

$$\zeta_j + G_j(X_{ji} - \zeta_j) = F_j + D_j(X_{ji} - F_j);$$

$$G_j^2 \sum_i \left(C_{ji} / \sum_h C_{jh} \right) (X_{ji} - X_{ji})^2 =$$

$$(1 - 1) D_j \sigma_j^2 / \sum_h C_{jh} + D_j^2 \sum_i \left(C_{ji} / \sum_h C_{jh} \right) (X_{ji} - X_{ji})^2, \quad (2.22)$$

where

$$X_{ji} = \sum_i X_{ji} C_{ji} / \sum_h C_{jh}. \quad (2.23)$$

2.3 Risques liés au redressement; paramètres de modèle connus par hypothèse

La méthode exposée dans la section précédente définit les niveaux de sous-dénombrement pour diverses combinaisons de strates et de régions comme des variables aléatoires. La perte *prévue* (ou *risque*) est le critère utilisé pour comparer des méthodes de redressement entre elles. En règle générale, on préfère les méthodes qui présentent le moins de risques.

où $f(C_i)$ est n'importe quelle fonction du chiffre de recensement de la région i . En minimisant (2.15) pour tous les $Y_{est}^i \equiv \sum_{j=1}^J F_{est}^{ji} C_{ji}^j$ nous sommes amenés à choisir les F_{est}^{ji} de telle manière que $E[\sum_{j=1}^J \lambda_{est}^{ji} (F_{est}^{ji} - F_{ji}^j) | \{X_{ji}: i = 1, \dots, I; j = 1, \dots, J\}]$ est minimisée; dans l'expression précédente, $\lambda_{ji} \geq 0$ dépend uniquement des chiffres du recensement $\{C_{ji}: i = 1, \dots, I; j = 1, \dots, J\}$. On obtient la valeur minimum de l'espérance mathématique par l'estimateur (2.14), qui révèle une certaine robustesse étant donné qu'il demeure optimal, peu importe la fonction $f(\cdot)$ choisie.

Conformément à la recommandation 7.2 de la National Academy of Sciences (1985), le choix de $f(C_i) = 1/C_i$ fait en sorte que la part de la perte totale pour une région est en rapport avec la taille de la population de cette région. Parmi les fonctions de perte qu'utilise le Census Bureau, celle qui se rapproche le plus de (2.15), où $f(C_i) = 1/C_i$, est

$$(2.16) \quad \sum_{i=1}^I (Y_{est}^i - Y_i)^2 / Y_i;$$

elle se rapproche le plus de (2.15) en ce sens qu'il s'agit aussi d'une somme de carrés pondérée, où chaque terme correspond à une part de la perte totale qui est en rapport avec la taille de la population. Dans le cas qui nous occupe, on attribue des poids plus élevés dans les régions plus peuplées de sorte que l'utilisation de fonctions de perte de ce genre dénote un intérêt marqué pour les considérations d'ordre national. La fonction de perte $\sum_{i=1}^I (Y_{est}^i - Y_i)^2 / Y_i^2$, qui garantit un taux de sous-dénombrement uniforme pour les I régions, ne sera pas considérée ici.

Il est facile de montrer que l'estimateur de Bayes pour la fonction de perte (2.16) est défini,

$$(2.17) \quad Y_{est}^i = \left[E \left(\left(\sum_{j=1}^J F_{ji} C_{ji}^j \right)^{-1} \mid \{X_{ji}: i = 1, \dots, I; j = 1, \dots, J\} \right) \right]^{-1},$$

qui n'est pas une combinaison linéaire de $\{F_{ub}^{ji}: j = 1, \dots, J\}$. Toutefois, en faisant une première approximation à l'aide de la méthode δ , on peut montrer que $Y_{est}^i \approx Y_{ub}^i$. De fait, cette relation est vraie pour une catégorie beaucoup plus grande de fonctions de perte, comme l'indique Cressie (1987b):

$$(2.18) \quad L_\lambda \equiv \frac{\lambda(\lambda + 1)}{2} \sum_{i=1}^I \left\{ Y_{est}^i \left[\left(\frac{Y_i}{Y_{est}^i} \right)^\lambda - 1 \right] + \lambda [Y_i - Y_{est}^i] \right\}; \lambda \neq 0, -1;$$

les cas $\lambda = 0$, -1 sont définis comme les limites respectives de L_λ lorsque $\lambda \rightarrow 0$, -1 . Read et Cressie (1988, Chapitre 8) montrent que l'estimateur de Bayes dans ce cas est

$$(2.19) \quad Y_{est}^i(\lambda) = \left[E \left(\left(\sum_{j=1}^J F_{ji} C_{ji}^j \right)^{-\lambda} \mid \{X_{ji}: i = 1, \dots, I; j = 1, \dots, J\} \right) \right]^{-1/\lambda},$$

qui se ramène à (2.14) lorsque $\lambda = -1$, et à (2.17) lorsque $\lambda = 1$. Les critères de la chose est que la plupart des estimateurs de taux de sous-dénombrement utilisés sont optimaux (suivant diverses hypothèses de modèle) pour $\lambda = -1$, bien que l'on utilise $\lambda = 1$; c'est-à-dire (2.16), pour mesurer leur efficacité. Par la méthode δ nous avons $Y_{est}^i(\lambda) \approx Y_{ub}^i$, et rappelons-nous que Y_{ub}^i est optimal pour (2.15); par conséquent, les estimateurs de taux de sous-dénombrement pour les fonctions de perte quadratiques sont efficaces pour une grande catégorie de fonctions de perte. C'est le cas observé par Kadane (1984) dans son analyse bayésienne hiérarchique des données du sous-dénombrement dans le recensement de 1980 ($\lambda = -1$ et $\lambda = -2$ ont été comparés) et ce qu'ont confirmé les études de populations fictives réalisées par Cressie et Dajani (1988).

2.2 Fonctions de perte (mesures de gain) et estimateurs de Bayes correspondants

La fonction de perte est utilisée en théorie de la décision statistique (voir p. ex. Ferguson 1967) pour quantifier la perte qui découle de l'utilisation de θ comme estimateur d'un paramètre alors que la valeur réelle est θ . Un exemple d'une terminologie plus optimiste, le Census Bureau a décidé en 1986 de parler de "mesure de gain" plutôt que de "fonction de perte".

Imaginons (2.10) comme une distribution de X_{ji} étant donné F_{ji} , (distribution conditionnelle) et (2.7) comme la distribution composée (ou "a priori") de F_{ji} . Pour prévoir F_{ji} nous devons alors connaître la distribution "a posteriori" de F_{ji} étant donné X_{ji} . Le lecteur aura remarqué que nous utilisons une terminologie bayésienne puisque nous considérons les F_{ji} comme des variables aléatoires dont la distribution empirique est conforme au modèle (2.7). Toutefois, outre ces paramètres aléatoires, il nous faut aussi estimer des paramètres fixes mais inconnus $\{F_j\}$, $\{\tau_j^2\}$, $\{\sigma_j^2\}$. La distribution a posteriori de $F_{ji} | X_{ji}$ est,

$$(2.11) \quad \frac{(\text{distribution de } X_{ji} | F_{ji}) \cdot (\text{distribution a priori de } F_{ji})}{(\text{distribution marginale de } X_{ji})}.$$

Pour ce qui a trait à la fonction de perte quadratique, l'estimateur de Bayes ordinaire de F_{ji} est simplement l'espérance de F_{ji} par rapport à la distribution a posteriori: $F_{ji}^{\text{uba}} = E(F_{ji} | X_{ji})$. En substituant les modèles (2.7) et (2.10) dans l'expression (2.11), nous obtenons facilement la distribution a posteriori (voir p. ex. Lindley et Smith 1972):

$$(2.12) \quad F_{ji} | X_{ji} \sim N \left(F_j + \frac{\tau_j^2}{\tau_j^2 + \sigma_j^2} (X_{ji} - F_j), \frac{\sigma_j^2 \tau_j^2}{\tau_j^2 + \sigma_j^2} / C_{ji} \right),$$

pour $i = 1, \dots, J$; $j = 1, \dots, I$. Par conséquent, l'espérance de F_{ji} par rapport à la distribution a posteriori est simplement

$$(2.13) \quad F_{ji}^{\text{uba}} = F_j + D_j (X_{ji} - F_j),$$

où $D_j \equiv \tau_j^2 / (\tau_j^2 + \sigma_j^2)$. Pour convertir (2.13) en un estimateur empirique de Bayes, il faut trouver des estimateurs pour F_j et D_j ; voir Section 3.1.

Bien que l'on se soit servi des hypothèses de normalité des modèles (2.7) et (2.10) pour établir l'équation (2.13), il est possible de montrer, de façon plus générale, que (2.13) est un estimateur de Bayes pour la fonction de perte quadratique en posant simplement par hypothèse la structure de la moyenne et de la variance des modèles (2.7) et (2.10) et en posant $E(F_{ji} | X_{ji}) = a_{ji} + b_{ji} X_{ji}$. Ceci représente un cas particulier de la solution plus générale présentée par Goldstein (1975). Pour des raisons de commodité, nous allons maintenant l'hypothèse de normalité mais il faut se rappeler qu'il existe une optimalité non paramétrique pour tous les estimateurs étudiés.

L'estimateur F_{ji}^{uba} défini en (2.13), est un estimateur de Bayes pour la fonction de perte quadratique dans la strate j de la région i . Définissons l'estimateur de Y_i ,

$$(2.14) \quad Y_i^{\text{uba}} \equiv \sum_{j=1}^J F_{ji}^{\text{uba}} C_{ji}; \quad i = 1, \dots, I,$$

et considérons la fonction de perte générale suivante :

$$(2.15) \quad \sum_{i=1}^I (Y_i^{\text{est}} - Y_i)^2 f(C_i),$$

L'objet de cet article n'est pas de supposer des F_{ji} qui ne dépendent que de j , ou une fonction de régression pour les F_{ji} mais de convertir l'hypothèse synthétique $F_{ji} = F_j$, en une hypothèse d'homogénéité (statistique) :

$$(2.7) \quad F_{ji} \sim N(F_j, \tau_j^2 / C_{ji}); \quad i = 1, \dots, I; j = 1, \dots, J,$$

où " \sim " signifie "est distribué selon" et $N(\mu, \sigma^2)$ est une distribution normale de moyenne μ et de variance σ^2 . En utilisant une fonction de régression pour la moyenne, on peut expliquer une plus grande partie de la variabilité des F_{ji} au risque d'introduire un biais par une mauvaise définition. Les strates qui ont été choisies dans la section 3 sont définies selon l'origine raciale; nous avons jugé qu'il était inutile de compliquer davantage la question en choisissant des variables de régression sujettes à controverse. Nous désignerons le modèle (2.7) comme une distribution composée. Pour des raisons de commodité, nous supposons au départ une *distribution normale; cette hypothèse sera assouplie ultérieurement*. Dans le modèle (2.7), F_j moyenne fixe mais inconnue qu'il faut estimer et $\tau_j^2 = \text{var}(\sqrt{C_{ji}} F_{ji})$ est un paramètre que nous appellerons *variance de strate* (uniformisée). Le modèle (2.7) reflète mieux la réalité aux niveaux d'agrégation supérieurs; voir Section 3. Par hypothèse, toutes les distributions dans (2.7) sont indépendantes.

Il y a de bonnes raisons de pondérer la variance par $1/C_{ji}$ (voir Cressie 1987a, Annexe; et 1988). L'aspect le plus intéressant du modèle (2.7) est qu'il est *cohérent*, c'est-à-dire qu'il ne change pas d'un niveau d'agrégation à l'autre. Plus précisément,

$$(2.8) \quad F_{j,ikl} \sim N\left(F_j, \frac{\tau_j^2}{C_{j,ikl}}\right),$$

où

$$(2.9) \quad F_{j,ikl} \equiv \frac{F_{ji} C_{ji} + F_{ji'} C_{ji'}}{F_{ji} C_{ji} + F_{ji'} C_{ji'}}, \text{ et } C_{j,ikl} \equiv C_{ji} + C_{ji'}.$$

Il s'agit d'une propriété très importante que la plupart des modèles statistiques de sous-dénombrement actuellement proposés *n'ont pas*. Grâce à cette propriété, les élaborateurs de modèles n'ont pas à s'occuper des facteurs géographiques et historiques qui ont contribué à diviser le pays en États, en comtés, etc. Évidemment, l'ensemble $\{F_{ji}; i = 1, \dots, I; j = 1, \dots, J\}$ n'est pas connu comme tel; s'il l'était, ce serait très facile de calculer, $\{Y_j; i = 1, \dots, I\}$. En fait, on procède à un échantillonnage, de sorte que F_{ji} est observé imparfaitement. Pour nous représenter le problème, imaginons qu'un échantillon est prélevé dans la strate j de la région i pour évaluer le niveau de sous-dénombrement. Soit X_{ji} le résultat (p. ex., X_{ji} est le rapport entre l'estimateur de système dual et les chiffres du recensement pour la strate j de la région i , et le modèle

$$(2.10) \quad X_{ji} \sim N(F_{ji}, \sigma_j^2 / C_{ji}); \quad i = 1, \dots, I; j = 1, \dots, J,$$

où F_{ji} est un paramètre de moyenne inconnue qu'il faut estimer et $\sigma_j^2 = \text{var}(\sqrt{C_{ji}} X_{ji})$ est un paramètre que nous appellerons *variance d'échantillonnage* (uniformisée). Par hypothèse, toutes les distributions dans (2.10) sont indépendantes. Lorsqu'il y a un grand nombre de strates, l'échantillon de l'EP doit être grand (p. ex. 300,000 ménages) si l'on veut obtenir des données pour chaque combinaison de strates et de régions. Dans ses enquêtes post-censitaires, le U.S. Census Bureau utilise l'échantillonnage avec probabilité proportionnelle à la taille, ce qui suppose une variance d'échantillonnage du type défini en (2.10). À cause de cette pondération, le modèle (2.10) est aussi cohérent.

Définitions

(2.1) Y_{ji} = population réelle dans la j -ième strate de la région i

(2.2) C_{ji} = population recensée dans la j -ième strate de la région i

(2.3) $F_{ji} \equiv Y_{ji}/C_{ji}; i = 1, \dots, I; j = 1, \dots, J.$

Supposons pour l'instant que nous connaissions les rapports $\{F_{ji}; j = 1, \dots, J\}$ pour la région i . Nous pouvons donc calculer la population réelle Y_i à l'aide des valeurs C_{ji} .

(2.4)
$$Y_i = \sum_j F_{ji} C_{ji}.$$

Les rapports F_{ji} sont souvent appelés *facteurs de redressement*. Les strates sont formées de telle manière que ces facteurs $\{F_{ji}; i = 1, \dots, I\}$ soient aussi homogènes que possible dans la strate $j; j = 1, \dots, J$ (Tukey 1981).

Dans la réalité, on ne connaît jamais les facteurs de redressement; des estimateurs synthétiques exploitent l'homogénéité et remplacent (2.4) de la façon suivante :

(2.5)
$$Y_{i\text{va}}^i = \sum_j F_j C_{ji}.$$

Nous n'avons plus que J facteurs de redressement synthétiques $\{F_j; j = 1, \dots, J\}$ à estimer, après quoi nous obtenons par (2.5) une estimation de Y_i . Les estimateurs synthétiques ont pour avantage de rendre les facteurs de redressement indépendants de i , ce qui nous permet de les appliquer à *n'importe quel* niveau d'agrégation.

Les facteurs de redressement (estimés) peuvent aussi faire l'objet de modèles de régression où les variables indépendantes peuvent ou non être des variables du recensement, par exemple, pourcentage de personnes faisant partie d'un groupe minoritaire, taux de criminalité, et pourcentage de personnes habituellement recensées. Considérons

(2.6)
$$Y_{i\text{reg}}^i = \sum_j \left(\sum_p \beta_{k,j} z_{k,j,i} \right) C_{ji}.$$

Pour ajuster efficacement les paramètres $\beta_{1,j}, \dots, \beta_{p,j}$, nous posons diverses hypothèses à propos des composantes d'erreur $\{F_{ji} - \sum_{k=1}^p \beta_{k,j} z_{k,j,i}\}$, notamment quelles sont indépendantes et identiquement distribuées avec une moyenne nulle.

Ericksen et Kadane (1985) proposent d'ajuster une droite de régression à $\sum_{j=1}^J F_{ji} C_{ji} / \sum_{j=1}^J C_{ji}; i = 1, \dots, I$. Freedman and Navidi (1986) désapprouvent l'idée et soulignent les conséquences qui pourraient se produire si l'une ou l'autre des hypothèses relatives à l'erreur s'avérait non fondée. Toutefois, ils semblent oublier que l'utilisation de ratios entraîne l'*héteroscédasticité* des erreurs, ce que nous ne manquons pas de souligner dans l'équation (2.7) ci-dessous; dans la section 2.2, nous justifions le choix de ce modèle. Par ailleurs, selon ce même modèle, les chiffres de sous-dénombrement des strates sont combinés de sorte que la variabilité entre les strates est expliquée par la fonction de régression et la variance de l'erreur. Il est possible d'obtenir des estimateurs plus précis par l'équation (2.6) en définissant une fonction de régression pour chaque strate. Ericksen et Kadane (1987) et Ericksen, Kadane et Tukey (1987) supposent aussi des erreurs homoscedastiques et un modèle de régression fondé sur la combinaison de strates hétérogènes. Il semble qu'une telle combinaison ait été rendue nécessaire par le manque de données.

du gouvernement fédéral puisqu'elles comptent proportionnellement plus de membres des groupes difficiles à dénombrer. De plus, certains Etats comme New York et la Californie ont le sentiment d'être sous-représentés au Congrès, au profit des Etats du Midwest comme l'Illiana et l'Iowa.

Le taux de sous-dénombrement est défini simplement comme la différence entre le chiffre réel et le chiffre du recensement, exprimée en pourcentage du chiffre réel. Pour estimer le taux de sous-dénombrement, nous allons nous servir d'un modèle fondé sur des données de l'enquête post-censitaire (EP). Cet article nous donnera l'occasion d'aborder certains aspects techniques d'une méthode de redressement fondée sur un modèle. Dans la section 2, nous définissons le modèle et nous nous intéressons au choix des mesures de gain; en outre, nous présentons les résultats d'aggrégation et de désaggrégation fondés sur des estimateurs de Bayes et des estimateurs synthétiques. Dans la section 3, nous présentons des résultats fondés cette fois sur des conditions suffisantes pour que le risque associé aux chiffres redressés soit moindre que le risque associé aux chiffres bruts.

2. LE MODELE COMPOSE ET SES CONSÉQUENCES

Nous aimerions tout d'abord expliquer la source de variation aléatoire dans notre modèle, lequel a été défini initialement dans Cressie (1986), puis amélioré dans Cressie (1988). Selon notre modèle, la population réelle de toute strate délimitée aux Etats-Unis est inconnue. Une fois que les chiffres du recensement correspondants sont connus, on met à jour les estimations de la population réelle. Autrement dit, toutes les inférences *reposent* sur les observations du recensement.

2.1 Le modèle

L'*estimation synthétique* consiste à estimer le taux de sous-dénombrement à un niveau parti-culier (par exemple l'Etat) en faisant la somme des niveaux de sous-dénombrement enregistrés dans les strates (p. ex., strates démographiques) de la région considérée (p. ex., la Californie); cette méthode d'estimation suppose que le rapport de la population réelle à la population recen-sée est *fixe* pour chaque strate, peu importe la région considérée (p. ex., le rapport pour les jeunes hommes de race noire est le même pour la Californie, le Delaware, etc.). Ces strates et le sexe. Tukey (1981) a toutefois proposé de former aussi des strates selon des facteurs géographiques et urbains. Isaki et coll. (1986) ont procédé à une telle stratification pour les Etats-Unis.

Le modèle composé que nous proposons ici suppose qu'une stratification a déjà été faite; malgré cela, nous proposons dans la section 4 une façon de déterminer *a posteriori* si le mode de stratification choisi est acceptable.

Supposons qu'il y a $J = 1, \dots, J$ strates et $i = 1, \dots, I$ régions (p. ex., lorsqu'il s'agit de secteurs de dénombrement, $I \approx 300,000$, tandis que lorsqu'il s'agit d'Etats, $I = 51$, y compris le District de Columbia; pour ce qui a trait à la stratification selon des facteurs démogra- phiques, $J = 30$ par exemple, tandis que pour ce qui a trait aux deux modes de stratification utilisés dans Isaki et coll. 1986, $J = 90$ et $I = 96$). Considérons que la strate j est fixe (p. ex., la strate j peut être constituée de la population noire vivant dans les principaux centres d'une zone statistique urbaine (SMSA = Standard Metropolitan Statistical Area) de 250,000 habi- tants ou plus dans la division de recensement de la Nouvelle-Angleterre. Comme i prend les valeurs de $1, \dots, I$, on obtient à chaque fois une série de sous-régions; la sous-région dési- gnée par l'indice " ji " représente la partie de la région i où se trouve la strate j . Nous ne consi- dérons que les sous-régions pour lesquelles il existe des *chiffres de recensement*.

Dans quelles circonstances les opérations de redressement amélioreraient-elles les chiffres du recensement?

NOEL CRESSIE¹

RÉSUMÉ

Des arguments convaincants militent pour ou contre le redressement des chiffres des recensements décennaux aux États-Unis mais bon nombre de ces arguments reposent plus sur des considérations politiques que techniques. La décision de redresser les chiffres du recensement dépend essentiellement de la méthode de redressement. De plus, si le redressement devait s'effectuer, par exemple, à l'aide d'une méthode synthétique ou d'une méthode de régression, à quel niveau devrait-il se faire et comment devrait-on procéder pour les niveaux inférieurs ou supérieurs? Pour apporter une réponse judicieuse à ces questions, il nous faut un modèle d'erreurs de sous-dénombrement "cohérent" en ce sens qu'il ne change pas d'un niveau d'agrégation à l'autre (pays, état, comté, etc.). Le présent article propose un modèle de ce genre; les sous-régions ayant des caractéristiques communes sont groupées par strate de telle sorte que les moyennes des facteurs de redressement des sous-régions de la strate soient les mêmes et que les variances soient inversement proportionnelles aux chiffres du recensement. En prenant en considération l'échantillonnage des régions (par l'estimation de système dual par exemple), nous pouvons construire des estimateurs empiriques de Bayes qui intègrent des éléments d'information sur la moyenne de la strate et la valeur de l'échantillon. Ces estimateurs sont calculés pour chaque état (51 états, y compris Washington, D.C.) et stratifiés selon l'origine raciale ou ethnique (3 strates) à l'aide de données de l'enquête post-censitaire de 1980 (PEP 3-8, pour la population hors établissement institutionnel).

MOTS CLÉS: Estimation empirique de Bayes; fonctions de perte; mesures de gain; fonction quantile; corrélation géographique; estimation synthétique.

1. INTRODUCTION

Cet article est de nature technique mais il importe, croyons-nous, d'exposer brièvement les dimensions politiques et sociales de la "question du sous-dénombrement" aux États-Unis. La loi oblige le U.S. Census Bureau à communiquer les chiffres de population des États au Congrès au plus tard le 31 décembre de l'année du recensement décennal en vue de la redistribution des sièges à la Chambre des Représentants et à communiquer, au plus tard le 31 mars 1991, les données démographiques régionales en vue de la redéfinition des districts. Les formes d'utilisation des données du recensement se sont multipliées au cours des dernières décennies: les formules de partage des revenus utilisent les chiffres de population et le revenu par habitant pour chaque localité constituée en municipalité; les études démographiques et sociologiques réalisées au niveau des régions, des états ou du pays en général reposent habituellement sur les chiffres du recensement, etc.

L'imprécision des chiffres du recensement devrait être un sujet de préoccupation pour tout le pays. Il ne fait pas de doute que certains groupes d'individus (par exemple, les jeunes Noirs, les immigrants illégaux, etc.) sont plus difficiles à dénombrer que d'autres; voir Erickson et Kadane (1985) et Freedman et Navidi (1986) de même que l'analyse qui vient à la suite de ces articles. Si les groupes difficiles à dénombrer étaient répartis également dans les régions politiques et administratives des États-Unis, la question du sous-dénombrement soulèverait beaucoup moins de controverse. À l'heure actuelle, plusieurs des grandes villes américaines comme Chicago, Détroit, New York et Los Angeles prétendent qu'elles devraient recevoir plus d'argent

¹ Noel Cressie, Département de statistique, Iowa State University, Ames, IA 50011.

De même, les questionnaires remis en retard ne doivent pas être traités comme les questionnaires qui ont été remis dans les délais prescrits. Comme la visite d'un intervieweur de l'EP peut avoir incité des ménages à remplir leur questionnaire du recensement, bien que tardivement, l'inclusion de ces questionnaires pourrait avoir pour conséquence d'introduire un biais dans l'estimation du taux de sous-dénombrement.

Dans le recensement de 1986, les «faux» questionnaires et les questionnaires remis en retard contenaient les noms de 115,000 personnes, soit 0.7% de la population. Ce nombre n'est pas inclus dans le chiffre du recensement brut (Y) ni dans le nombre estimé de personnes (selon l'EP) qui ont été recensées (y), mais il l'est dans le nombre estimé de personnes (selon l'EP) qui auraient dû être recensées (x). Autrement dit, les personnes dont le nom figurait dans de «faux» questionnaires ou des questionnaires remis en retard ont été considérées comme oubliées et le redressement pertinent a été effectué à l'aide de (x). Le facteur de redressement (x/y) est gonflé du fait que les «faux» questionnaires et les questionnaires remis en retard sont exclus de (y); en revanche, les deux séries de questionnaires sont aussi exclus du chiffre du recensement brut (Y), ce qui compense.

Méthode d'estimation

L'estimation s'est faite selon l'âge, le sexe et la région géographique (division statistique de la capitale et de l'Etat). Des facteurs de redressement ont été inclus dans les formules d'estimation pour tenir compte partiellement des ménages qui n'avaient pas répondu ou qui n'avaient pu être contactés. Ces facteurs permettent de redresser les deux principales valeurs estimées, x et y, en imputant effectivement, pour chaque ménage qui a refusé de répondre ou qui n'a pas été contacté, le nombre moyen de membres d'un ménage et, pour chaque personne ainsi imputée, le taux de sous-dénombrement moyen pour la combinaison âge x sexe x région géographique appropriée. Pour réduire le biais attribuable à l'utilisation de tels facteurs, ceux-ci ont été calculés pour divers sous-groupes de ménages selon le code de dénombrement attribué au recensement (par exemple logement occupé, questionnaire remis en retard). Ce code de dénombrement était réputé avoir un rapport avec le taux de non-réponse observé dans l'EP.

BIBLIOGRAPHIE

- BAILAR, B.A. (1985). Comments on "Estimating the population in a Census Year: 1980 and beyond" par E.P. Ericksen et J.B. Kadane, *Journal of the American Statistical Association*, 80, 109-114.
- BISHOP, Y.M.M., FIENBERG, S.A., et HOLLAND, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge: MIT Press.
- FAY, R., PASSEL, J.S., et ROBINSON, J.G. (1988). The coverage of population in the 1980 Census. Evaluation and Research Report, PHC 80-E4, United States Bureau of the Census, Washington D.C.
- PURCELL, N.J., et KISH, L. (1979). Estimation for small domains. *Biometrics*, 35, 365-384.
- WOLTER, K.M. (1986). Capture-recapture estimation in the presence of known sex ratio. SRD Research Report, United States Bureau of the Census, Washington, D.C.

Lorsque l'adresse du ménage était imprécise, par exemple dans les régions rurales, les préposés du traitement devaient utiliser aussi le nom du membre responsable du ménage, le nom de la propriété, etc. Pour mieux identifier le ménage. De plus, les préposés devaient passer en revue toutes les adresses inscrites dans le registre de manière à déceler les enregistrements répétés. Ils vérifiaient aussi les adresses figurant dans les registres des DC voisins lorsque l'adresse du ménage échantillonné était aux limites du DC.

La deuxième étape consistait à apparter les enregistrements et on se servait pour cela du nom et des caractéristiques démographiques des personnes enregistrées lors du recensement et de l'EP. Durant le processus d'appariement, une formule de recherche était produite chaque fois que l'adresse indiquée pour un membre du ménage échantillonné différait de l'adresse du logement. La formule de recherche était traitée comme la formule d'interview de l'EP et on tentait de repérer le questionnaire du recensement qui correspondait à l'adresse indiquée sur la formule de recherche.

En règle générale, le processus d'appariement se déroulait sans difficulté. Dans certains cas toutefois, des erreurs d'orthographe et des adresses incomplètes rendaient plus difficile l'appariement de noms. Pour déterminer alors si la personne en question avait été recensée ou non, on se fondait sur d'autres renseignements comme l'âge, le sexe, l'état matrimonial, le lieu de naissance et le lien avec les autres membres du ménage. Dans l'incertitude, les préposés du traitement devaient consulter leur superviseur.

Durant l'interview, le répondant devait aussi indiquer si le nom de chaque membre du ménage figurait dans un questionnaire du recensement. Lorsque l'appariement échouait, faute de renseignements pertinents, la déclaration du répondant était néanmoins tenue pour vraie. Il y a eu quelques cas où ce genre de renseignements n'ont pas été fournis par le répondant; on a alors considéré les personnes en question comme n'ayant pas été recensées. Après l'appariement, les données ont été enregistrées sur bande, puis contrôlées et restructurées de manière à obtenir un fichier d'enregistrements unitaires épuré, qui indique le nombre de fois qu'une personne incluse dans l'échantillon de l'EP a été enregistrée au cours du même recensement.

Traitement des questionnaires du recensement remis en retard et des «faux» questionnaires

En définissant la formule d'estimation

$$X = Y(x/y), \text{ où}$$

$$X = \text{chiffre du recensement estimé, redressé en fonction du sous-dénombrement}$$

$$Y = \text{chiffre brut du recensement, non redressé}$$

$$x = \text{nombre estimé de personnes (selon l'EP) qui auraient dû être recensées}$$

$$y = \text{nombre estimé de personnes (selon l'EP) qui ont été recensées,}$$

on a considéré les «faux» questionnaires du recensement et les questionnaires remis en retard comme des cas d'omission dans le recensement.

De «faux» questionnaires du recensement ont été créés sur le terrain pour les logements que l'on savait habités mais dont les ménages occupants n'avaient pas retourné leur questionnaire du recensement et n'avaient pu être rejoints. Les agents recenseurs devaient exercer beaucoup de prudence en créant ces «faux» questionnaires et devaient s'assurer, hors de tout doute, que les logements en question étaient occupés le soir du recensement. Ils devaient aussi recueillir autant de renseignements que possible sur le nombre de membres de ces ménages et leurs caractéristiques démographiques.

Lorsqu'on réussissait à apparter une adresse relevée durant l'EP à un «faux» questionnaire du recensement, il était impossible d'exécuter l'appariement de façon satisfaisante à cause de l'absence de noms et de caractéristiques personnelles fiables dans le questionnaire du recensement.

Comme les personnes demeurant en permanence dans des logements non privés ou spéciaux et celles vivant dans les régions peu peuplées représentent globalement moins de 3% de la population totale, une différence de taux de sous-dénombrement entre ces groupes et les groupes inclus dans le champ de l'EP a peu de chances d'influer beaucoup sur le taux de sous-dénombrement estimé pour l'Etat ou le pays en général.

Lien entre le recensement et l'EP

Il est essentiel que l'EP soit aussi indépendante que possible du recensement, sinon les facteurs qui ont favorisé un sous-dénombrement ou un surdénombrement lors du recensement peuvent se retrouver dans l'EP, ce qui aurait pour effet d'introduire un biais dans l'estimation du taux de sous-dénombrement. De plus, le fait de connaître les régions ou secteurs qui doivent être inclus dans le champ de l'EP pourrait conditionner le comportement des agents recenseurs dans ces secteurs, de sorte que l'échantillon de l'EP ne refléterait pas fidèlement le taux de sous-dénombrement dans le recensement. C'est pourquoi on a tenu à mobiliser deux équipes distinctes pour le recensement et l'EP, tant sur le terrain qu'au bureau central. Les intervieweurs de l'EP n'avaient pas agi comme agents recenseurs ou chefs d'équipe et le personnel affecté aux opérations sur le terrain ne connaissait pas les régions ou secteurs qui allaient être inclus dans le champ de l'EP.

L'indépendance des enquêtes était assurée de deux autres façons : par la création de deux systèmes de collecte qui fonctionnent indépendamment l'un de l'autre et par l'adoption de procédures spéciales concernant les questionnaires du recensement reçus par la poste après que les opérations de l'EP ont été engagées.

Pour assurer l'indépendance des opérations des deux enquêtes, on a débuté l'EP après que tous les questionnaires du recensement dûment remplis ont été ramassés. Ainsi, les agents recenseurs n'étaient pas sur le terrain au même moment que les intervieweurs de l'EP et il n'y avait aucune possibilité qu'ils s'influencent les uns les autres, même involontairement.

Il fallait prévoir des procédures spéciales pour les questionnaires du recensement reçus après le début de l'EP pour annuler l'effet que pouvaient avoir les opérations de l'EP sur les membres responsables du ménage, qui n'avaient pas remis leur questionnaire à temps. À certaines adresses, les intervieweurs de l'EP ont découvert des questionnaires du recensement qui n'avaient pas encore été ramassés. Cela s'expliquait par le fait que les personnes en question préféreraient retourner leur questionnaire par la poste et ne l'avaient pas encore fait ou que l'agent recenseur n'avait pu entrer en contact avec ces personnes pour ramasser le questionnaire. Si ce n'avait été de l'EP, quelques-unes de ces personnes n'auraient peut-être jamais retourné leur questionnaire du recensement. Pour éliminer tout risque de biais attribuable à ce comportement, les questionnaires du recensement retournés par la poste après le lundi, 20 juillet 1986 (jour où a débuté l'EP) étaient considérés en retard. Des procédures spéciales pour le traitement des questionnaires remis en retard sont énoncées plus loin dans cette annexe.

Procédures d'appariement de l'EP

Les opérations d'appariement visant à déterminer si une personne a été oubliée, recensée une seule fois ou recensée plus d'une fois, se sont déroulées en deux étapes. Celles-ci consistaient en des procédés administratifs appliqués par les employés du Centre de transcription des données du recensement.

La première étape consistait à repérer les questionnaires du recensement correspondant aux adresses des ménages inclus dans l'échantillon de l'EP. Les opérations de traitement du recensement de 1986 étaient centralisées à Sydney. Les employés du Centre de transcription des données du recensement devaient comparer l'adresse indiquée sur la page frontispice de la formule d'interview de l'EP avec toutes les adresses qui figuraient dans le registre de l'agent recenseur responsable du district de collecte (DC) où demeurait le ménage en question. Le registre précité était un moyen de contrôle pour la livraison et le ramassage des questionnaires du recensement et contenait des renseignements comme le nom et l'adresse des membres de tous les ménages du DC de même que le nombre de personnes dans chacun de ces ménages.

REMERCIEMENT

Nous tenons à remercier les personnes qui ont révisé cet article pour leurs commentaires et leurs propositions concernant des sujets de recherche.

ANNEXE 1

L'ENQUÊTE POSTCENSITAIRE DE 1986

Généralités

L'EP de 1986 a été réalisée dans la quatrième et cinquième semaine qui ont suivi le jour du recensement. À cet occasion, on a interviewé un échantillon de la population vivant dans environ 35,000 logements privés (2/3 de un pour cent des logements) répartis dans toute l'Australie et abritant environ 100,000 personnes. Le taux de sondage variait selon l'Etat ou le Territoire, les Etats ou Territoires moins étendus ayant des taux de sondage plus élevés. Les intervieweurs recueillaient des données personnelles comme le nom, l'âge, le sexe, l'état matrimonial et le lieu de naissance en vue de les appairer avec les données contenues dans les questionnaires du recensement. Pour chaque personne visée par l'enquête, l'intervieweur inscrivait le lieu de résidence habituel, le lieu où la personne se trouvait le soir du recensement, l'adresse de la personne avant et après le jour du recensement et toute autre adresse à laquelle la personne aurait pu être recensée. À chaque adresse donnée, les renseignements personnels étaient comparés aux données contenues dans les questionnaires du recensement afin de déterminer si la personne en question avait été recensée ou non ou de définir le nombre de fois qu'elle avait été recensée si elle l'avait été plus d'une fois.

Champ de l'EP et structure de l'échantillon

Hormis les cas particuliers mentionnés ci-dessous, le champ de l'EP comprenait toutes les personnes qui devaient avoir été recensées, sauf celles qui étaient à l'étranger ou qui sont décédées entre le jour du recensement et le jour de l'EP. Les agents diplomatiques et les personnes en résidence diplomatique n'ont pas été recensés; ces gens ont donc été exclus du champ de l'enquête, comme les enfants nés après le recensement. Pour les personnes qui étaient à l'étranger le jour du recensement mais qui étaient incluses dans le champ de l'EP, on a tenté un appariement avec les questionnaires du recensement pour voir si elles n'avaient pas été recensées par erreur.

Pour des raisons pratiques, les régions très peu peuplées ont été exclues du champ de l'enquête. Lors du recensement, des procédures spéciales avaient été appliquées pour rejoindre et recenser la population aborigène, les travailleurs des camps miniers et des fermes d'élevage, etc. Pour réaliser l'EP dans ces régions, il aurait fallu appliquer les mêmes procédures qu'au recensement, ce qui n'aurait pas permis un calcul juste et indépendant du taux de sous-dénombrément. C'est pourquoi ces régions ont été exclues du champ de l'EP.

Les logements non privés ou spéciaux comme les hôpitaux, les hôtels et les motels ont aussi été exclus du champ de l'enquête. La grande majorité des personnes qui se trouvent dans ces logements y sont habituellement pour une période relativement courte et selon les règles normales du ABS en matière d'enquête, ces personnes seraient susceptibles d'être enregistrées à leur lieu de résidence habituel, où seraient fournis les renseignements pertinents. Par cette règle, on s'est donc trouvé à exclure du champ de l'enquête un nombre relativement restreint de personnes qui demeurent en permanence dans ces logements. Pour les besoins de l'estimation, on a supposé que le taux de sous-dénombrément pour la population ne faisant pas partie du champ d'enquête était le taux moyen pour la capitale ou une autre ville, selon le cas, pour chaque Etat et le taux moyen pour le Territoire dans le cas des deux Territoires.

Wolter considère deux modèles. Dans le premier, il suppose que le degré de corrélation entre les probabilités de sous-dénombrement et dans l'EP (lequel degré est mesuré par les rapports de produits croisés établis à l'aide de tables comme le diagramme exposé dans les pages précédentes) est le même pour les hommes et les femmes dans chaque groupe d'âge. Dans le second modèle, Wolter pose l'hypothèse d'indépendance pour les femmes et utilise un rapport de masculinité déterminé à l'aide de données indépendantes pour calculer le degré de corrélation pour les hommes. Il est alors possible de déduire les rapports de produits croisés pour les hommes.

Selon une première étude, où l'on a appliqué ces méthodes à des données de l'Australie, on a observé que le premier modèle produisait des estimations de rapports de produits croisés très irrégulières, la moitié environ des valeurs estimées étant négatives. Le second modèle produisait beaucoup moins de valeurs estimées négatives mais il en restait un certain nombre; ces valeurs négatives allaient être ramenées à zéro dans un modèle modifié. Wolter (1986, p. 7) souligne aussi le problème de produits croisés négatifs. Une fois modifié, le second modèle a été appliqué à des données de 1986. Les rapports de masculinité tirés des données de l'EP pour les groupes d'âge 5-9 à 35-39 étaient conformes aux prévisions et leur application donnait exactement la valeur estimée de l'EP. En ce qui a trait au groupe d'âge 0-4, on a utilisé un rapport de masculinité établi à l'aide d'estimations démographiques et pour les groupes d'âge 40-44 et suivants, on s'est servi de rapports de masculinité estimés à l'aide des chiffres du recensement. Les rapports de masculinité sont reproduits dans le tableau 6.

Comme le rapport de masculinité utilisé n'est pas très différent des rapports de masculinité de l'EP, l'application du second modèle de Wolter a pour effet de modifier très légèrement les estimations de l'EP. Pour ce qui est des groupes d'âge 0-4 et 75+, le nombre estimé de personnes de sexe masculin augmente de 0,7% et de 0,5% respectivement. En ce qui a trait aux groupes d'âge 45-49 et 70-74, les estimations sont réduites d'environ 0,6%. Cette analyse donne à penser que l'action combinée des problèmes évoqués ci-dessus créerait des différences de biais relativement faibles entre les sexes dans la méthode d'estimation de l'EP. Il se pourrait que des biais soient à peu près identiques pour les hommes et pour les femmes, de sorte que les rapports de masculinité de l'EP seraient généralement acceptables.

Les recensements de 1981 et de 1986 ont prouvé la nécessité des rapports de masculinité dans l'évaluation des mesures de sous-dénombrement et nous croyons que la méthode de Wolter est un moyen utile de produire des estimations auxquelles peuvent être comparés les chiffres du recensement et les estimations de l'EP. Le fait que les rapports de masculinité de l'EP de 1986 ont été généralement acceptables signifie que l'utilisation de la méthode de Wolter a changé peu de choses. Surtout en ce qui concerne le groupe d'âge 0-4, la situation est tout à fait différente de celle qu'on a connue en 1981, lorsqu'on a jugé nécessaire de redresser les estimations de l'EP pour certains groupes d'âge en se fondant sur des rapports de masculinité indépendants. Cette différence de situation entre 1981 et 1986 reflète peut-être une diminution du degré de corrélation entre les probabilités de sous-dénombrement dans le recensement et dans l'EP en 1986.

8. CONCLUSION

Tandis que l'ABS a redressé les chiffres des trois derniers recensements pour tenir compte du sous-dénombrement, la fiabilité des estimations de l'EP tient au fait que cette enquête produit des résultats comparables à ceux d'autres sources. On ne prévoit pas une modification majeure de la méthode pour le recensement de 1991. Cependant, nous croyons nécessaire d'approfondir les causes possibles de biais en vérifiant notamment la conformité des méthodes d'appariement manuelles et en analysant les méthodes qui visent à éliminer le biais de corrélation. On prévoit aussi d'étudier la possibilité de mettre sur pied une banque de données démographiques selon le lieu de résidence habituel de manière à éliminer ou à réduire les effets des nombreux mouvements migratoires à court terme.

Fois. Elle a aussi permis de découvrir que des personnes avaient été recensées par erreur. De cette façon, si on assimile le processus d'estimation de l'EP à l'estimation par quotient plutôt qu'à l'estimation de système dual, on est plus en mesure d'établir le nombre d'enregistrements erronés.

L'estimation de système dual suppose que le fait qu'une personne soit oubliée ou non dans l'EP n'a rien à voir avec le fait qu'elle ait été recensée ou non. Bien que l'on ait pris toutes les mesures nécessaires pour séparer parfaitement les deux systèmes de collecte et de traitement, il peut toujours exister une certaine corrélation. Une corrélation positive signifiera que la valeur estimée de l'EP fondée sur l'hypothèse d'indépendance donne une sous-estimation; une corrélation négative signifiera que la valeur estimée de l'EP donne une surestimation. Il y a corrélation négative si le fait qu'une personne a été recensée implique qu'il sera difficile de la dénombrer dans l'EP; cependant, il n'y a pas de donnée qui permette d'affirmer cela sans contredit; le taux de réponse final pour l'EP (95%) est conforme à ceux observés dans d'autres enquêtes-ménages de l'ABS. Une corrélation positive semble plus probable; du reste, l'existence d'une telle corrélation aurait été constatée dans le recensement de 1981. Si, de fait, il y a corrélation positive, les opérations de redressement fondées sur l'EP n'auront pas donné des résultats tout à fait satisfaisants mais auront été néanmoins valables.

7. AUTRES MÉTHODES D'ESTIMATION (WOLTER 1986)

La combinaison de données de l'EP et de rapports de masculinité établis à l'aide de données démographiques ou d'autres sources est à la base des méthodes proposées par Wolter (1986). Wolter propose en effet plusieurs modèles et méthodes connexes qui combinent formellement des rapports de masculinité avec des estimations de l'EP. Ces méthodes visent à assouplir l'hypothèse d'indépendance qui est caractéristique des méthodes d'estimation de l'EP.

Tableau 6
Rapports de masculinité: Hommes pour 100 femmes

Âge	Autre	EP
0-4	105.0	104.3
5-9	105.2	105.2
10-14	105.2	105.3
15-19	104.7	104.7
20-24	104.1	104.1
25-29	102.6	102.6
30-34	100.3	100.3
35-39	102.4	102.4
40-44	104.5	104.9
45-49	105.2	106.0
50-54	104.2	104.7
55-59	103.0	103.5
60-64	95.2	95.2
65-69	87.1	87.2
70-74	78.8	79.2
75+	57.9	57.6

exagérément élevées, par rapport aux corrections qu'il faut faire pour tenir compte du sous-dénombrement. On a donc procédé de façon indirecte pour estimer la population des États et des Territoires selon l'âge et le sexe en appliquant une méthode d'ajustement proportionnel itératif (API) aux estimations de l'EP se rattachant à un niveau supérieur et comportant une faible erreur d'échantillonnage. Pour avoir une description de l'API, voir Purcell et Kish (1979). L'API consiste à redresser les chiffres du recensement selon l'âge et le sexe en fonction des estimations nationales de la population selon l'âge et le sexe et des estimations de la population des États ou Territoires selon le sexe.

L'API comporte les cycles suivants $n = 0, 1, \dots$

$$X_{(2n+1)}^{gas} = X_{(2n)}^{gas} \frac{X_{as}^{(2n)}}{X_{as}^{(2n+1)}} \\ X_{(2n+2)}^{gas} = X_{(2n+1)}^{gas} \frac{X_{gs}^{(2n+1)}}{X_{gs}^{(2n+2)}}$$

et $X_{(0)}^{gas} = Y^{gas}$ le chiffre du recensement pour l'État g , le groupe d'âge a et le sexe s . Cette méthode converge vers une solution unique. L'utilisation de l'API suppose évidemment que la relation entre les variables dans la structure d'association est valide et qu'elle est préservée. En ce qui concerne les estimations de population pour les districts d'administrations locales, la question de l'erreur d'échantillonnage est plus délicate et les résultats de l'EP ne sont pas assez fiables pour que l'on puisse estimer directement le taux de sous-dénombrement qui s'applique à chaque district d'administration locale. Compte tenu de l'hypothèse que le taux de sous-dénombrement varie selon l'âge et le sexe ou le lieu de naissance (né en Australie ou né à l'étranger) et qu'il diffère suivant l'État ou Territoire et selon qu'il s'agit de la capitale ou du reste de l'État, le redressement des estimations touchant les districts d'administrations locales a été exécuté de manière à tenir compte de tous ces critères.

6. PROBLÈMES LIÉS AU PROCESSUS D'ESTIMATION DE L'EP

Comme le souligne Bailar (1985) par exemple, le biais et la convergence des estimations de l'EP sont influencés par les erreurs d'appariement, par l'existence d'une corrélation entre le fait qu'une personne a été oubliée dans le recensement et le fait qu'elle a été aussi oubliée dans l'EP et par l'existence d'enregistrements erronés dans le recensement ou l'EP. C'est précisé-ment à cause des effets probables de ces facteurs que l'on évalue les résultats de l'EP en utilisant de la façon décrite ci-dessus des données démographiques et administratives.

Les erreurs d'appariement auront pour effet d'introduire un biais dans les estimations de l'EP. Le non-appariement d'enregistrements qui devraient normalement pouvoir être appariés aura pour effet de gonfler le nombre de personnes oubliées dans le recensement et l'EP pro-duira alors des valeurs surestimées. Les cas de "fausse concordance" auront l'effet contraire. La présence d'enregistrements erronés dans le recensement ou l'EP aura pour effet de gonfler la valeur de X ou de x et, par voie de conséquence, l'estimation établie par l'EP. Le U.S. Bureau of the Census prélève un échantillon spécial (appelé échantillon D) parmi les enregistrements du recensement afin d'estimer le nombre d'enregistrements erronés dans le recensement; on suppose l'incidence de ces enregistrements sur les estimations officielles en redressant le chiffre du recensement X . Pour une description de l'échantillon D, voir Fay, Passel et Robinson (1988). Les méthodes d'appariement et d'estimation utilisées par l'ABS visent à corriger en partie l'effet des enregistrements erronés en vérifiant non seulement si une personne a été recensée mais encore si elle aurait dû l'être ou si elle l'a été plus d'une fois. Par exemple, l'EP de 1986 a permis de déterminer que 250 personnes avaient été recensées deux fois et 4 personnes l'avaient été trois

Le second problème a trait au taux de sous-dénombrement estimé très élevé qui a été établi par l'EP pour le Territoire du Nord. Comme l'indique le tableau 4, ce taux est de 9,97% lorsqu'on tient compte du lieu de résidence habituel. Le Territoire du Nord est un région peu peuplée (selon le recensement de 1986, 154,800 habitants répartis sur 1,3 million de kilomètres carrés) qui est caractérisée par une population très mobile. La population estimée du Territoire du Nord selon l'EP est considérablement plus élevée que la population estimée selon le recensement de 1981. En comparant les estimations de l'EP pour le Territoire du Nord avec des estimations indépendantes inscrites à l'école, on se rend compte également que les estimations de l'EP sont élevées. Même en reconnaissant que les estimations indépendantes peuvent être entachées d'erreur, nous pouvons affirmer avec assez de certitude que l'EP a surestimé le taux de sous-dénombrement pour le TN.

Après avoir vérifié les questionnaires de l'EP pour le Territoire du Nord, on a jugé qu'ils étaient acceptables sauf pour un district de collecte, où des problèmes de contrôle des adresses et de déplacement dans des régions peu accessibles ont fait ressortir des lacunes dans les instructions concernant les opérations sur le terrain et ont engendré des problèmes d'appariement. On a jugé bon de rajuster à la baisse le nombre estimé de filles de 0 à 4 ans et la population estimée du TN en redressant les résultats de l'EP. En ce qui a trait aux filles de 0 à 4 ans, on a multiplié le rapport de masculinité établi à l'aide des estimations démographiques par le nombre estimé de garçons de 0 à 4 ans selon l'EP. Cela revenait essentiellement à remplacer l'estimation établie par l'EP pour les filles de 0 à 4 ans par une meilleure estimation en ayant recours au rapport de masculinité et au nombre estimé de garçons selon l'EP. On a pu ainsi réduire de 4,000 le nombre estimé de filles de 0 à 4 ans, le nouvel effectif estimé étant de 587,000. En ce qui concerne la population du TN, on a décidé de ne pas se servir des données du district de collecte qui posait problème. Cela a eu pour effet de ramener le taux de sous-dénombrement pour le Territoire du Nord à 9,1% (calculé selon le lieu de résidence effectif) et à 5,5% (selon le lieu de résidence habituel).

Par suite de ce double redressement, le taux de sous-dénombrement pour l'ensemble du pays est passé de 1,91% à 1,87% (selon le lieu de résidence effectif) ou de 1,84% à 1,81% (selon le lieu de résidence habituel). Le tableau 5 donne les estimations de l'EP selon l'âge et le sexe après que les estimations relatives à la population du TN et aux filles de 0 à 4 ans ont été redressées.

5. ESTIMATIONS DÉMOGRAPHIQUES INFRA-NATIONALES

Les données sur la migration interne ne sont pas assez fiables pour que les estimations démographiques infra-nationales servent à évaluer le taux de sous-dénombrement dans le recensement. Néanmoins, on a comparé pour chaque Etat et chaque Territoire le nombre estimé d'enfants de 1 à 15 ans selon l'EP de 1986 avec le nombre d'enfants inscrits au registre des allocations familiales. De façon générale, on observe une concordance entre les deux séries de données sauf pour ce qui a trait au Territoire du Nord, où l'écart est supérieur à 2%. Compte tenu de cette concordance entre les deux séries de données et du fait que l'on ne peut comparer les estimations de l'EP à des données indépendantes fiables touchant les groupes d'âge supérieurs à cause de l'absence de telles données, les estimations (redressées) de l'EP concernant la population des Etats et des Territoires ont été acceptées.

Les estimations de la population des Etats et des Territoires selon l'âge et le sexe et de la population des districts d'administrations locales n'ont pas été tirées directement de l'EP. L'EP de 1986 était une enquête par sondage et ses résultats sont exposés à l'erreur d'échantillonnage. Lorsqu'il s'agit d'estimations selon l'âge et le sexe au niveau de l'Etat ou du Territoire ou encore au niveau du district d'administration locale, les erreurs d'échantillonnage sont élevées, souvent

Tableau 5
Chiffres du recensement de 1986 redressés en fonction du sous-dénombrement,
selon l'âge et le sexe

Age	Selon le lieu de résidence effectif			
	Hommes		Femmes	
	Nombre (milliers)	Taux de sous- dénombrement	Nombre (milliers)	Taux de sous- dénombrement

Tous les âges	Selon le lieu de résidence habituel			
	Hommes		Femmes	
Age	Nombre (milliers)	Taux de sous- dénombrement	Nombre (milliers)	Taux de sous- dénombrement
	Total		Total	
0-4	616.3	1.30	586.6	1.17
5-9	602.4	1.24	572.4	1.27
10-14	670.1	1.39	636.8	1.38
15-19	688.3	2.19	657.3	2.02
20-24	679.4	4.54	652.4	2.95
25-29	677.5	4.17	660.7	1.81
30-34	629.9	2.29	627.8	1.55
35-39	634.0	1.87	618.9	1.11
40-44	512.6	1.64	488.5	1.21
45-49	426.9	1.66	403.0	0.98
50-54	371.2	2.04	354.6	1.56
55-59	379.5	1.62	366.5	1.06
60-64	347.0	1.70	364.4	1.70
65-69	263.6	1.52	302.3	1.35
70-74	208.2	1.92	262.9	1.47
75+	233.0	1.49	404.7	2.08
Tous les âges	7940.1	2.16	7959.7	1.58

Age	Selon le lieu de résidence habituel			
	Hommes		Femmes	
Tous les âges	Nombre (milliers)	Taux de sous- dénombrement	Nombre (milliers)	Taux de sous- dénombrement
	Total		Total	
0-4	615.3	1.29	585.9	1.22
5-9	601.3	1.23	571.2	1.22
10-14	668.5	1.29	635.7	1.36
15-19	685.6	2.11	654.3	1.97
20-24	673.1	4.33	646.9	2.83
25-29	672.6	4.02	657.2	1.80
30-34	626.6	2.21	625.6	1.53
35-39	630.9	1.78	616.7	1.05
40-44	510.3	1.59	487.0	1.19
45-49	424.7	1.52	401.7	0.98
50-54	369.6	1.97	353.0	1.52
55-59	377.7	1.52	364.0	0.92
60-64	345.6	1.74	361.6	1.61
65-69	262.1	1.47	300.2	1.31
70-74	207.2	1.89	261.3	1.46
75+	232.4	1.52	403.3	2.01
Tous les âges	7903.6	2.08	7925.5	1.54

1.26	1201.2
1.23	1172.5
1.33	1304.2
2.04	1339.9
3.59	1320.0
2.92	1329.8
1.87	1252.2
1.41	1247.6
1.39	997.3
1.26	826.4
1.75	722.6
1.22	741.8
1.67	707.3
1.38	562.3
1.65	468.5
1.83	635.7
1.81	15829.1

les garçons du même âge et les filles de 5 à 9 ans). L'EP a permis d'estimer à 11,300 le nombre de filles de 0 à 4 ans qui ont été oubliées dans le recensement, comparativement à environ 7,000 pour les filles de 5 à 9 ans. Cette différence sensible n'existe pas du côté des garçons.

Le rapport de masculinité établi par l'EP pour le groupe d'âge 0-4 est de 104.3 garçons pour 100 filles, ce qui est inférieur au rapport de masculinité établi à l'aide des chiffres du recensement (104.9) et au rapport de masculinité établi à l'aide des données démographiques (105.0 garçons pour 100 filles).

Les chiffres ci-dessus donnent à penser que l'EP a surestimé le nombre de filles de 0 à 4 ans bien que nous voyions difficilement comment cette enquête aurait pu surestimer l'effectif de ce groupe plutôt que l'effectif de tout autre groupe.

Tableau 3
 Comparaison entre les résultats de l'EP de 1986 et des estimations indépendantes

Estimations de l'EP	Estimations démographiques	Allocations familiales	Inscriptions scolaires
0- 4	1207.3	1196.6	1204.8(a)
5- 9	1174.8	1168.5	1177.0
10-14	1307.2	1298.6	1304.2
			1289.6
			-(b)

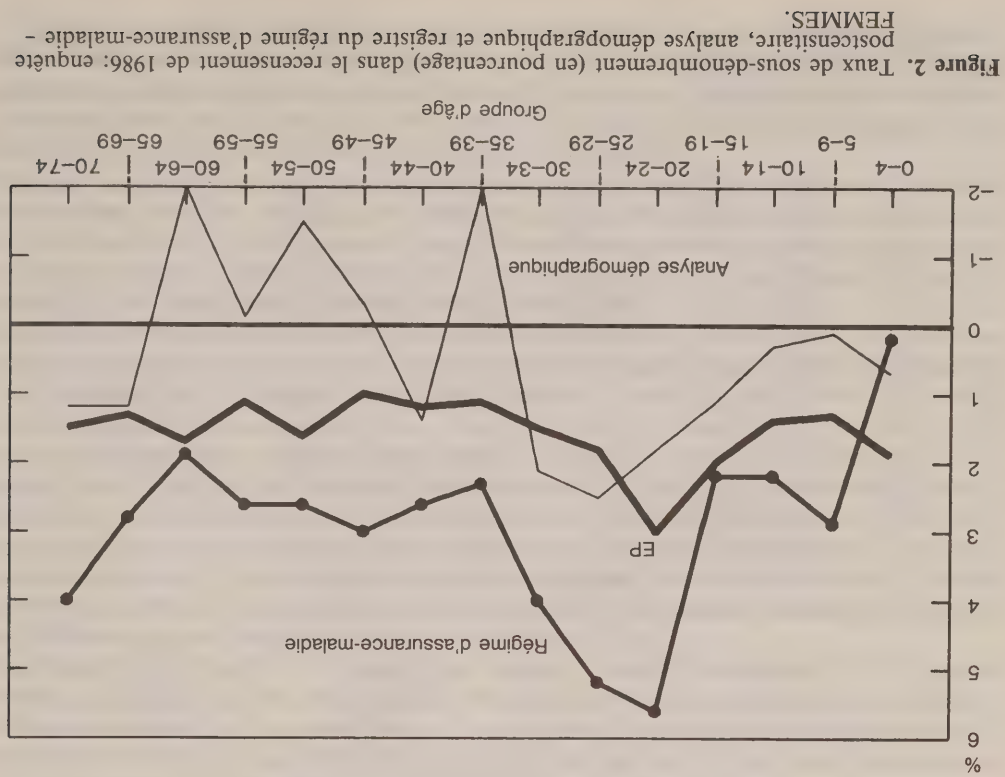
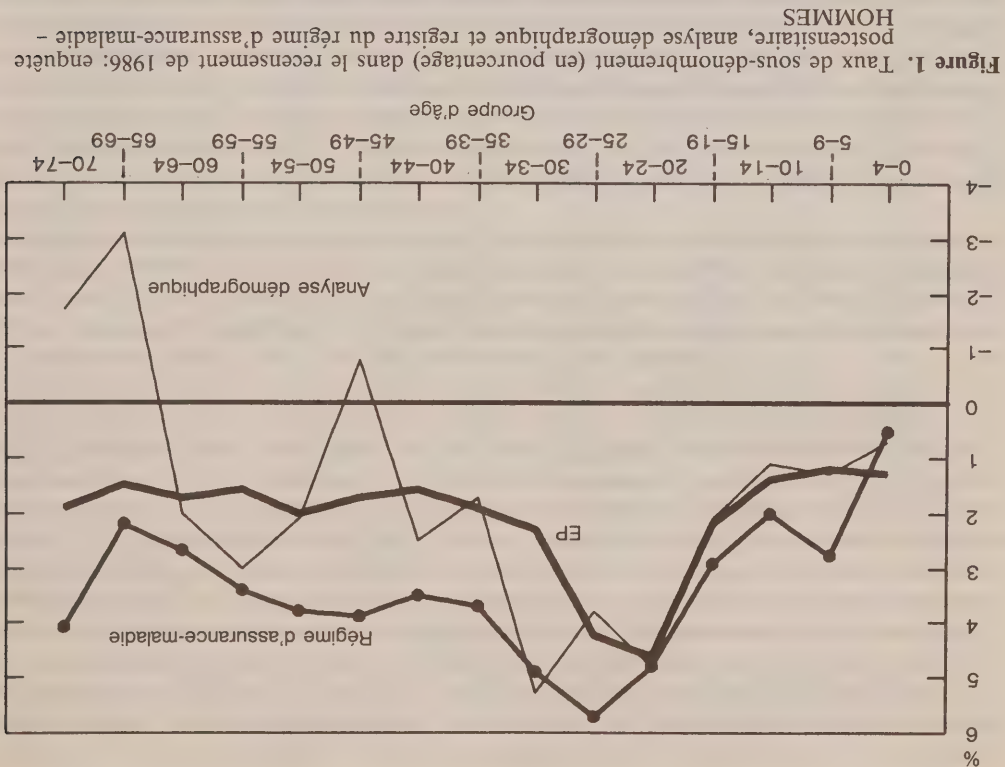
Personnes (milliers)

(a) Le nombre d'enfants d'âge 0 inscrits au programme d'allocations familiales est sous-estimé à cause du délai qui s'écoule entre le jour de la naissance et le moment où l'enfant est inscrit au programme d'allocations familiales. Pour redresser le chiffre en question, on a remplacé le nombre d'enfants d'âge 0 inscrits au programme par une estimation tirée de l'analyse démographique.

(b) Les enfants de 5 ans ne sont pas obligés de fréquenter l'école.

Tableau 4
 Taux de sous-dénombrement (%) selon l'état ou le territoire (étabi à l'aide des données de l'EP)

Suivant le lieu de résidence habituel	Suivant le lieu de résidence effectif	
1.51	1.54	Nouvelles-Galles-du-Sud
1.77	1.59	Victoria
2.43	2.68	Queensland
1.59	1.54	Australie méridionale
2.26	2.32	Australie occidentale
1.16	1.32	Tasmanie
6.45	9.97	Territoire du Nord
1.61	1.95	Territoire fédéral de la capital
1.84	1.91	Australie



Il existe une très forte correspondance entre les estimations de l'EP et les estimations démographiques concernant la population de sexe masculin, plus particulièrement en ce qui a trait aux hommes de moins de 30 ans. On note toutefois un écart appréciable en ce qui concerne les hommes de 30 à 34 ans, les estimations démographiques excédant de 20,000 les estimations de l'EP. Cet écart peut trouver son explication dans les mouvements migratoires à court terme qui ont eu lieu en Australie entre 1981 et 1986 et qui ont amené une forte augmentation nette du nombre d'hommes appartenant à ce groupe d'âge. Comme on n'observe pas d'augmentation de cet ordre dans les groupes d'âge voisins, on pourrait croire qu'il s'est glissé des erreurs dans les données sur les courants migratoires. Les mouvements migratoires étant très nom- breux en Australie (plus de 6 millions en 1986), une petite erreur dans la déclaration de l'âge ou dans le traitement peut fausser dans une assez grande mesure les estimations démographiques en chiffres absolus nets. La possibilité d'erreurs dans les estimations démographiques est d'autant plus réelle que le taux de sous-dénombrement pour le groupe des 30 à 34 ans (5.3%) est très élevé par rapport aux taux pour les groupes d'âge voisins.

De fait, il est très possible que l'EP ait mal mesuré le taux de sous-dénombrement des visiteurs d'outre-mer. Quoi qu'il en soit, les erreurs d'estimation concernant la portion de la popula- tion qui est de passage en Australie ne devraient pas influencer sur le degré de précision des estima- tions officielles de la population puisque celles-ci sont fondées sur le principe du lieu de résidence habituel et ne comprennent pas les visiteurs.

En ce qui concerne les femmes, le degré de correspondance entre les résultats de l'EP et les estimations démographiques pour les personnes de moins de 35 ans est satisfaisant. Toutefois pour certains groupes d'âge, les estimations démographiques sont beaucoup moins élevées que les estimations de l'EP, et pour les personnes de 35 à 39 ans et de 45 à 64 ans, les estimations démographiques paraissent franchement trop faibles, ce qui tend à confirmer que les estimations démographiques ne sont pas suffisamment précises pour produire des estima- tions de la population et devraient servir uniquement à évaluer les résultats des EP.

Les taux de sous-dénombrement établis à l'aide des données de l'EP pour les divers groupes d'âge sont beaucoup plus uniformes que les taux établis à l'aide de l'analyse démographique. Les taux établis par l'EP pour les personnes de 20 à 29 ans sont plus élevés que pour les per- sonnes des autres groupes d'âge et cela est normal étant donné les taux de mobilité plus élevés chez les jeunes adultes, particulièrement les hommes.

Le nombre de personnes inscrites au régime d'assurance-maladie est beaucoup plus élevé que les estimations de l'EP et les estimations démographiques, sauf en ce qui concerne le groupe des 0 à 4 ans, où c'est le contraire. Selon certaines études, cet écart s'explique par le délai qui s'écoule entre la naissance d'un enfant et le moment où celui-ci est inscrit au régime d'assurance- maladie; de même, pour ce qui est des autres groupes d'âge, l'écart s'explique par le délai qui s'écoule entre le moment où une personne meurt ou émigre et le moment où son nom est rayé du registre de l'assurance-maladie.

En comparant les estimations de l'EP à des estimations tirées du registre des allocations familiales et du registre des inscriptions scolaires pour certains groupes d'âge, on constate un degré de correspondance acceptable. Ces résultats tendent à confirmer la précision des estima- tions de l'EP en ce qui concerne les premiers groupes d'âge.

Même s'il existe un degré de correspondance acceptable entre les estimations de l'EP et d'autres estimations de la population, il subsiste deux problèmes que nous devons tenter de résoudre avant de pouvoir accepter les estimations de l'EP. Le premier problème vient d'une analyse des estimations de l'EP concernant les taux de sous-dénombrement dans le recense- ment selon l'âge et le sexe. Ces taux figurent dans le tableau 2.

Sauf pour les groupes de 0 à 4 ans et de 75 ans et plus, les taux de sous-dénombrement sont généralement plus élevés pour les hommes que pour les femmes. Tandis que les taux de sous- dénombrement pour les personnes de 75 ans et plus peuvent être influencés par une faible taille d'échantillon, le taux pour les filles de 0 à 4 ans semble exagéré (1.9% par rapport à 1.3% pour

Tableau 2

Chiffres estimés de la population de 1986 selon l'âge et le sexe (fondés sur les résultats de l'EP de 1986 et de l'analyse démographique) et nombre de personnes inscrites au régime d'assurance-maladie

Hommes (milliers)									
Population					Écart par rapport au recensement			Taux de sous-dénombrement (en pourcentage)	
Recensement					Régime d'assurance-maladie			Régime d'assurance-maladie	
Age	EP(a)	AD(b)	Régime d'assurance-maladie	EP	AD	EP	AD	EP	AD
0-4	608.3	612.8	611.4	8.0	4.5	3.1	1.3	0.7	0.5
5-9	594.9	603.0	612.3	7.5	8.1	17.4	1.2	1.3	2.8
10-14	660.8	670.4	668.4	674.3	9.6	7.6	13.5	1.4	2.0
15-19	673.1	688.4	687.7	693.1	15.3	14.6	20.0	2.2	2.9
20-24	648.5	679.5	681.3	681.3	31.0	32.8	32.8	4.6	4.8
25-29	649.2	677.7	675.0	688.5	28.5	25.8	39.3	4.2	5.7
30-34	615.5	630.0	650.1	647.5	14.5	34.6	32.0	2.3	4.9
35-39	622.2	634.2	632.7	646.2	12.0	10.5	24.0	1.9	3.7
40-44	504.2	512.6	517.0	522.3	8.4	12.8	18.1	1.6	3.5
45-49	419.8	427.0	416.5	436.8	7.2	7.7	17.0	1.7	3.9
50-54	363.7	371.2	371.4	377.9	7.5	7.7	14.2	2.0	3.8
55-59	373.4	379.5	384.9	386.6	6.1	11.5	13.2	1.6	3.4
60-64	341.1	347.0	348.1	350.6	5.9	7.0	9.5	1.7	2.7
65-69	259.6	263.6	251.8	265.5	4.0	-7.8	5.9	1.5	2.2
70-74	204.2	208.2	200.8	213.0	4.0	-3.4	8.8	1.9	4.1
75+	229.5	233.0	181.5	250.1	3.5	-48.0	20.6	1.5	8.2
Total	7768.3	7941.0	7883.1	8057.3	172.7	114.8	289.0	2.2	3.6

Femmes (milliers)									
Population					Écart par rapport au recensement			Taux de sous-dénombrement (en pourcentage)	
Recensement					Régime d'assurance-maladie			Régime d'assurance-maladie	
Age	EP(a)	AD(b)	Régime d'assurance-maladie	EP	AD	EP	AD	EP	AD
0-4	579.7	591.0	583.8	580.9	11.3	4.1	1.2	1.9	0.2
5-9	565.1	572.4	565.5	582.1	7.3	0.4	17.0	1.3	2.9
10-14	628.0	636.8	630.2	641.8	8.8	2.2	13.8	1.4	2.2
15-19	644.1	657.4	651.4	666.3	13.3	7.3	22.2	2.0	3.3
20-24	633.1	652.5	644.4	670.4	19.4	11.3	37.3	3.0	5.6
25-29	648.7	660.7	665.4	684.1	12.0	16.7	35.4	1.8	5.2
30-34	618.1	627.8	631.2	643.9	9.7	13.1	25.8	1.5	4.0
35-39	612.1	619.1	600.2	626.3	7.0	-11.9	14.2	1.1	2.3
40-44	482.6	489.6	495.4	495.4	6.0	7.0	12.8	1.2	2.6
45-49	399.1	403.0	397.9	411.6	3.9	-1.2	12.5	1.0	3.0
50-54	349.1	354.6	343.9	358.6	5.5	-5.2	9.5	1.6	2.6
55-59	362.6	366.5	362.4	372.4	3.9	-0.2	9.8	1.1	2.6
60-64	358.2	364.4	351.3	365.3	6.2	-6.9	7.1	1.7	1.9
65-69	298.2	302.2	301.9	306.7	4.0	3.7	8.5	1.3	2.8
70-74	259.0	262.9	262.2	269.7	3.9	3.2	10.7	1.5	4.0
75+	396.2	404.7	385.0	434.1	8.5	-11.2	37.9	2.1	8.7
Total	7833.8	7964.6	7866.2	8109.6	130.8	32.4	275.8	1.6	3.4

(a) Selon le lieu effectif de résidence
(b) Estimations démographiques fondées sur le recensement de la population de 1921 et des événements démographiques survenus par la suite.

3. ESTIMATION DU SOUS-DÉNOMBREMENT DANS LE RECENSEMENT À L'AIDE DE DONNÉES DÉMOGRAPHIQUES

On peut aussi estimer le taux de sous-dénombrement dans le recensement au moyen de données démographiques comme celles tirées de recensements antérieurs et des registres des naissances et des décès de même que les données sur la migration externe. Par exemple, on peut estimer la population à une date précise en mettant à jour les données d'un recensement antérieur à l'aide des données sur les naissances, les décès et la migration externe. Plus le recensement en question est éloigné dans le temps, plus la série chronologique des données de l'état civil et des données sur la migration nécessaires est étendue et moins on a se fier sur la justesse des données du recensement de référence, car le nombre estimé de personnes nées après la date du recensement pertine ne tient qu'à la fiabilité des données sur les naissances, les décès et la migration. Les données sur la migration interne en Australie ne sont pas suffisamment fiables pour que l'on puisse utiliser des méthodes démographiques dans l'estimation de taux de sous-dénombrement au niveau infra-national. Par conséquent, seuls les totaux nationaux peuvent servir à évaluer la qualité d'un recensement.

Les données sur les naissances et les décès en Australie existent sous forme de séries chronologiques qui remontent jusqu'au 19^e siècle et ces séries sont vraisemblablement complètes. Dans chacun de ses rapports publiés de 1911 à 1961 après chaque recensement de la population, le statisticien du Commonwealth australien affirmait que la très grande majorité des naissances et des décès avaient été consignés mais reconnaissait qu'il pouvait y avoir eu quelques omissions et qu'il y avait des retards dans les inscriptions. Le rapport du statisticien n'est plus publié depuis 1961. Cependant, il n'y a rien qui indique que les naissances et les décès sont enregistrés avec moins de régularité qu'auparavant.

L'Australie tient aussi depuis longtemps des statistiques complètes et fiables sur les courants migratoires. Ces statistiques portent sur tous les mouvements migratoires, qu'il s'agisse des mouvements permanents ou des mouvements à court terme et à long terme. Cependant, elles présentent plusieurs lacunes qui limitent leur utilité en ce qui a trait à l'évaluation de la qualité des recensements. Premièrement, il y a des périodes dans l'histoire de l'Australie où, croit-on, les courants migratoires n'auraient pas été enregistrés correctement (par exemple, durant la Seconde Guerre mondiale et la période qui a suivi immédiatement cette guerre). En deuxième lieu, comme les mouvements migratoires à court terme sont à la hausse depuis 1960, on n'utilise plus qu'un échantillon d'enregistrements (environ 1 sur 20) pour l'analyse statistique depuis 1971. En troisième lieu, il peut arriver que les voyageurs soient classés dans la mauvaise catégorie de mouvements migratoires. Pour éviter ces erreurs de classification, on compare les estimations démographiques, les chiffres du recensement et les estimations de l'EP pour le jour du recensement en fonction du lieu effectif de résidence, qui englobe les trois catégories de mouvements migratoires.

Pour évaluer le taux de sous-dénombrement dans le recensement de 1986, nous avons établi des estimations de la population selon l'âge et le sexe en date du recensement en nous servant de données sur les naissances, les décès et la migration externe qui remontent à 1921 ainsi que des résultats du recensement de 1921. Les estimations de la population de 65 ans et moins reposent donc essentiellement sur les données relatives aux naissances, aux décès et à la migration et ne sont aucunement influencées par le degré de précision du recensement de 1921.

4. VALIDATION DES ESTIMATIONS DE L'EP DE 1986

Le tableau ci-dessous donne les chiffres estimés de la population au 30 juin 1986 selon l'âge et le sexe; ces chiffres sont fondés sur les résultats de l'analyse démographique et de l'EP de 1986. Le tableau donne également le nombre de personnes inscrites au régime d'assurance-maladie selon l'âge et le sexe.

à la fois des cas de surdénombrement et des cas de sous-dénombrement. À cet égard, la méthode utilisée diffère de la méthode de saisie-resaisie classique.

Pour estimer la variance, nous avons considéré X comme une estimation par quotient tirée d'un échantillon à plusieurs degrés. Les erreurs types relatives des estimations de l'EP figurent dans le tableau 1. D'après ce tableau et les tableaux 2 et 4, nous pouvons voir que les erreurs types sont beaucoup moins élevées que les corrections découlant des estimations nationales de l'EP selon l'âge et le sexe et selon l'Etat et le sexe.

Pour une description plus détaillée de l'enquête postcensitaire de 1986 et des méthodes d'estimation, voir l'annexe.

Tableau 1

Recensement de 1986: Erreurs types relatives des estimations démographiques de l'EP

Age	Hommes	Femmes	Total
%	%	%	%
0-4	0.29	0.36	0.24
5-9	0.29	0.30	0.22
10-14	0.28	0.29	0.21
15-19	0.32	0.32	0.24
20-24	0.49	0.43	0.34
25-29	0.49	0.36	0.32
30-34	0.39	0.34	0.27
35-39	0.36	0.30	0.24
40-44	0.38	0.32	0.26
45-49	0.37	0.30	0.25
50-54	0.43	0.38	0.30
55-59	0.38	0.30	0.25
60-64	0.41	0.38	0.29
65-69	0.43	0.37	0.29
70-74	0.53	0.41	0.34
75+	0.47	0.39	0.31
Tous les âges	0.12	0.10	0.08
Etat	Hommes	Femmes	Total
%	%	%	%
Nouvelles-Galles-du-Sud	0.21	0.18	0.14
Victoria	0.23	0.21	0.16
Queensland	0.27	0.24	0.19
Australie méridionale	0.27	0.20	0.17
Australie occidentale	0.29	0.25	0.19
Tasmanie	0.36	0.31	0.25
Territoire du Nord	1.65	1.53	1.22
Territoire fédéral de la capitale	0.61	0.74	0.55

Après pondération, on multiplie ce facteur de redressement par le chiffre du recensement (Y) pour obtenir une estimation de la population (X), c'est-à-dire $X = Y (x/y)$.
 Pour tenir compte des écarts entre le rendement prévu et le rendement réel de l'échantillon dans l'EP, on applique cette méthode en fonction de l'âge (groupes de cinq années), du sexe et de la région géographique (capitale/reste de l'Etat). Les estimations de l'EP sont produites aussi bien selon le lieu de résidence le jour du recensement que selon le lieu de résidence habituel. Le processus d'estimation comporte également un facteur de redressement pour les cas de non-contact et de non-réponse dans l'EP, qui sont peu nombreux. Par exemple, l'estimation de la population (selon le lieu de résidence habituel) pour la région géographique (s) et la cellule âge-sexe (a) est:

$$X_{sa} = Y_{sa} x_{sa}/y_{sa}$$

où

$$x_{sa} = \sum_{gc} \frac{D_{gc}}{D_{gc} + d_{gc}} \cdot \frac{f_g}{x_{sagc}}$$

et

$$y_{sa} = \sum_{gc} \frac{D_{gc}}{D_{gc} + d_{gc}} \cdot \frac{f_g}{y_{sagc}}$$

Dans ces formules d'estimation, l'indice c désigne le code de dénombrement du ménage de l'EP pour le recensement et l'indice g désigne la région géographique où a été échantillonnée la personne pour les besoins de l'EP. D_{gc} représente le nombre de ménages répondants et d_{gc} est le nombre de ménages de la région g et de la classe de dénombrement (c) qui n'ont pu être contactés ou n'ont pas répondu à l'enquête. Le taux de sondage varie d'un Etat à l'autre et est désigné par f_g .
 Défini de cette façon, l'estimateur par quotient stratifié a posteriori. Si nous faisons abstraction pour l'instant du fait que des personnes peuvent être recensées incorrectement ou plus d'une fois, l'estimateur en question est celui que l'on obtient par un système dual ou une méthode de saisie-résaisie dont font état notamment Bishop, Fienberg et Holland (1975, p. 231-234). C'est ce qu'illustre le diagramme ci-dessous, où, en vertu de l'hypothèse d'indépendance, l'estimation de la population totale est $Y (x/y)$ qui correspond à l'estimation par quotient X.

EP

Recensement	Inclus	Inclus	Exclus
	Inclus	Y	Y
	Exclus		
		X	

Or, l'EP de 1986 visait à faire connaître le nombre de personnes qui avaient été oubliées dans le recensement et le nombre de personnes qui avaient été recensées en trop, c'est-à-dire celles qui avaient été recensées par erreur ou plus d'une fois. La valeur estimée X tient compte

recensement de 1976 et des recensements subséquents n'a pas soulevé d'opposition et en aucun temps on n'a mis en doute le bien fondé d'une telle opération ou la précision des méthodes utilisées. C'est tout le contraire de ce qui a été vécu aux Etats-Unis lorsqu'on a voulu redresser les chiffres du recensement de 1980 pour tenir compte du sous-dénombrement; cette opération avait alors soulevé une grande controverse.

Pour évaluer le niveau de sous-dénombrement, on se sert surtout des données d'une EP. On évalue les résultats d'une telle enquête en les comparant à des estimations établies à partir de données démographiques et d'autres données indépendantes comme celles concernant les inscriptions scolaires, les enfants dont les parents reçoivent des allocations familiales et les personnes inscrites au régime d'assurance-maladie de l'Etat. En Australie, les enfants de 6 à 15 ans sont obligés de fréquenter l'école et toutes les mères qui ont un enfant de moins de 17 ans reçoivent des allocations familiales; du moins, il en était ainsi avant que l'on mette en application le test de moyennes, en novembre 1987. Enfin, le régime d'assurance-maladie est obligatoire et universel. Ces sources indépendantes de données sont donc utiles pour la vérification des résultats de l'EP et des estimations démographiques.

Les estimations de la population sont les seules données du recensement qui sont redressées pour tenir compte du sous-dénombrement. Les chiffres du recensement sont publiés sans avoir été redressés.

2. L'ENQUÊTE POSTCENSITAIRE DE 1986

Pour son recensement quinquennal de la population, l'Australian Bureau of Statistics (ABS) emploie des agents recenseurs pour livrer les questionnaires à chaque ménage et les ramasser une fois qu'ils sont remplis. Les personnes sont recensées en fonction de l'endroit où elles se trouvent le soir de la journée du recensement.

Grâce à ce réseau d'agents recenseurs, l'étape de la collecte des données ne s'étend pas au-delà de deux semaines suivant la date du recensement. On peut ainsi réaliser une EP quelque temps seulement après la date du recensement – en 1986, moins de 4 ou 5 semaines après le jour du recensement. Comme l'EP renferme un certain nombre de questions qui exigent des réponses détaillées sur l'endroit où se trouvait une personne le soir du recensement, le fait qu'elle soit réalisée peu de temps après la date du recensement réduit au minimum les erreurs de mémoire et amoindrit aussi le nombre d'exclusions attribuables à des décès ou à des voyages à l'étranger.

Puisque l'EP produit les données qui servent à redresser les chiffres du recensement pour tenir compte du sous-dénombrement, il est essentiel que cette enquête soit statistiquement indépendante du recensement. Dans l'annexe nous décrivons comment on assure l'indépendance des deux enquêtes.

Dans l'EP de 1986, la méthode utilisée consistait essentiellement à prélever un échantillon de personnes indépendamment du recensement par un échantillonnage aréolaire à plusieurs degrés de logements privés. Des membres de l'équipe régulière d'intervieweurs du ABS ont été envoyés dans les ménages échantillonnés pour recueillir les données pertinentes sur chaque membre du ménage; sur place, les renseignements devaient être fournis par une personne adulte responsable. Dans une troisième étape, le personnel du Centre de transcription des données du recensement a tenté d'apparier les enrégistrement du recensement et ceux de l'EP pour déterminer si chaque personne de l'échantillon devait avoir été recensée et combien de fois son nom figurait, le cas échéant, sur la liste des personnes recensées. Les méthodes utilisées sont décrites dans l'annexe.

Les résultats de l'EP permettent d'estimer le rapport entre le nombre de personnes qui auraient dû être recensées (x) et le nombre de personnes qui, selon les estimations, l'ont été réellement (y). Ce rapport est le facteur de redressement net, qui tient compte à la fois du surdénombrement et du sous-dénombrement des personnes.

Redressement des chiffres du recensement de 1986 en Australie pour le sous-dénombrement

C.Y. CHOI, D.G. STEEL et T.J. SKINNER¹

RÉSUMÉ

En Australie, les estimations démographiques sont établies à partir des chiffres du recensement; lors des trois derniers recensements (1976, 1981 et 1986), ces chiffres ont été redressés pour tenir compte du sous-dénombrement. L'opération de redressement s'inspire des résultats d'une enquête postcensitaire et d'une analyse démographique. Cet article expose les méthodes utilisées de même que les résultats du redressement des données de 1986. Les auteurs voient aussi dans l'usage formel des rapports de masculinité proposé par Wolter (1986) une amélioration par rapport au rôle qu'avaient auparavant ces ratios dans le redressement des chiffres du recensement.

MOTS CLÉS: Sous-dénombrement dans le recensement; enquête postcensitaire; estimations démographiques; rapports de masculinité.

1. INTRODUCTION

Le recensement de la population fournit les données de base qui permettent d'établir des estimations de la population pour le pays en général, chacun des huit États et les districts d'administrations locales. En Australie, ces estimations démographiques sont nécessaires pour déterminer le nombre de sièges dont disposera chaque État à la Chambre des représentants du parlement fédéral, le mode de répartition des fonds fédéraux entre les États et le niveau de financement des administrations locales. Les estimations démographiques servent aussi, comme il se doit, d'indices de la croissance et de la répartition démographiques et de dénominateurs dans divers indicateurs démographiques, sociaux et économiques. Compte tenu de l'importance de ces usages, les estimations démographiques doivent être d'un haut degré de précision.

En Australie, on reconnaît que le niveau de sous-dénombrement dans le recensement est élevé et qu'il a trait à d'importantes variables telles que le lieu de naissance, la région géographique et l'âge-sexe. C'est pourquoi les chiffres du recensement qui servent à l'établissement d'estimations démographiques sont redressés en fonction du sous-dénombrement.

Le redressement des chiffres du recensement en fonction du sous-dénombrement est une pratique récente en Australie. Avant le recensement de 1976, on se servait de chiffres non redressés pour estimer la population. On s'est rendu compte de la nécessité de redresser les chiffres du recensement lorsqu'on a constaté que les chiffres de 1976 étaient bien en-deçà des estimations démographiques qui avaient été établies pour 1976 par suite d'une mise à jour des données du recensement de 1971, et que l'enquête postcensitaire (EP) de 1976 révélait un taux de sous-dénombrement de 2,6%, comparativement à 0,5% en 1966 et à 1,3% en 1971. L'EP de 1976 révélait aussi des différences de taux de sous-dénombrement notables entre les États et les Territoires, les taux allant de 4,2% pour le Territoire du Nord, à 1,1% pour la Tasmanie. En 1986, le taux de sous-dénombrement a été estimé à 1,9%. Comme en 1976, on a relevé des différences notables entre les États et les Territoires. Le redressement des chiffres du

¹ C.Y. Choi, D.G. Steel et T.J. Skinner, Australian Bureau of Statistics, P.O. Box 10, Belconnen, ACT, 2616, Australie.

- STATISTIQUE CANADA (1987). Méthodes d'estimation de la population, Canada. N° 91-528F au catalogue, Statistique Canada.
- STATISTIQUE CANADA (1988). Taux de sous-dénombrement provenant de la contre-vérification des dossiers de 1986. Bulletin d'information à l'intention des utilisateurs, n° 2, Statistique Canada.
- STOTO, M.A. (1987). Statement to the Subcommittee on Census and Population, Committee on Post Office and Civil Service, U.S. House of Representatives, San Francisco.
- WILK, M.B. (1981). Lettre à M. H. Breaux, Président, Groupe d'étude parlementaire sur les accords fiscaux fédéraux-provinciaux, le 3 juillet 1981.

Ceux-ci soulèvent certaines questions fondamentales quant à la philosophie et à la politique qui doivent régir le fonctionnement d'un programme statistique, mettant en évidence la question connexe de l'ajustement pour le sous-dénombrement, thème sous-jacent de ce texte. Le modèle fondé sur le recensement met l'accent sur la stabilité et la cohérence interne du système d'estimation. Quant à lui, le modèle indépendant du recensement opte pour la souplesse de manière à atteindre la précision maximale en recourant à toutes les données pertinentes disponibles, même si ce devait être au prix d'un affaiblissement de la cohérence méthodologique dans le temps. La solution de ce dilemme dépendra dans une large mesure des succès remportés dans les quatre domaines d'activité statistique susmentionnés.

REMERCIEMENTS

L'auteur tient à exprimer sa gratitude pour tous les entretiens enrichissants qu'il a pu avoir avec ses collègues de la Division de la démographie: Gwenael Cartier, Céline Fortier, Gilbert Lagrange, Ronald Raby, Robert Riordan, Edward Shin et Ravi Verma. Il voudrait aussi remercier K.G. Basavarajappa, David Binder, Malcolm Britton, Dick Carter, Ivan Fellegi, Yolande Lavoie et M.P. Singh, pour leurs précieux commentaires.

BIBLIOGRAPHIE

- CARTER, R.G. (1988). Measuring coverage errors in the census population. Communication présentée à la réunion annuelle de la Canadian Population Society tenue du 4 au 7 juin 1988 à l'Université de Windsor en Ontario.
- COALE, A.J. (1955). The population of the United States in 1950 classified by age, sex and color. *Journal of the American Statistical Association*, 50, 16, 54.
- FAY, R.E., PASSEL, J.S., ROBINSON, G.J., et CONRAN, C.D. (1988). The coverage of population in the 1980 Census, United States Department of Commerce, Bureau of the Census.
- FELLEGI, I.P. (1980). Les chiffres du recensement dans l'affectation des fonds devraient-ils être corrigés? Quelques considérations d'ordre statistique. Conférence sur le sous-dénombrement au recensement: Procès-verbal de la conférence de 1980 (en anglais). United States Department of Commerce, Bureau of the Census, 193-203.
- FREEDMAN, D.A., et NAVIDI, W.C. (1986). Regression models for adjusting the 1980 Census. *Statistical Science*, 1, 3-39.
- KEYFITZ, N. (1979). Information and allocation: Two uses of the 1980 Census. *The American Statistician*, 33, 45-50.
- KEYFITZ, N. (1980). Issues in adjusting for the 1980 Census undercount. Mémoire présenté au congrès annuel de l'American Statistical Association, Détroit.
- KISH, L. (1980). Diverse Adjustments for Missing Data. *Proceedings of the 1980 Conference on Census undercount*, United States Bureau of the Census, 193-203.
- LAPIERRE-ADAMCZYK, E. (1970). Estimation, à l'aide des techniques d'analyse démographique, du sous-dénombrement net au recensement suivant l'âge et le sexe. Document de travail, Section de l'analyse démographique et de la recherche, Statistique Canada
- ROMANUC, A., et RABY, R. (1980). Impact du sous-dénombrement au recensement sur certains accords de partage des recettes entre l'administration fédérale et les provinces. Division de la démographie, Statistique Canada.
- SPENCER, B. (1980). Issues of accuracy and equity in adjusting for census undercoverage. Document présenté au congrès annuel de l'American Statistical Association, Détroit.

consécutifs, ceux de 1971, 1976 et 1981. Aussi, même si le chiffre du recensement était légèrement inférieur au chiffre «réel» de la population, il constituait une base particulièrement fiable pour mesurer l'accroissement de la population.

Toutefois, les résultats du recensement de 1986 s'écartent de cette tendance, le taux de sous-dénombrement estimé par la contre-vérification des dossiers s'élevant à 3,2%. Si on accepte la méthode des composantes comme critère de validation, le recensement de 1986 sous-évalue d'environ 20% l'accroissement de la population au cours de la période de 1981 à 1986. La contre-vérification des dossiers et l'analyse démographique confirment donc toutes deux la détérioration de la couverture du recensement en 1986.

Pour ce qui est des composantes de l'accroissement démographique, l'autre fondement du système d'estimation, l'enregistrement des naissances, des décès et des immigrants reçus est passablement fiable. Comme nous l'avons vu à la section précédente, l'estimation de la migration interprovinciale et celle de l'émigration ont bénéficié, pour leur part, de diverses améliorations des méthodes et des données, particulièrement depuis 1981. Toutefois, rappelons que ces estimations reflètent certaines faiblesses inhérentes à des sources qui, comme les fichiers des allocations familiales et de l'impôt sur le revenu, ont été conçues à des fins administratives plutôt que statistiques. Elles demeurent, avec le sous-dénombrement et le surdénombrement au recensement, les principales sources d'erreurs et de biais dans les estimations postcensitaires de la population par province.

Que réserve l'avenir au système d'estimation que nous venons de décrire? Doit-il persister tel qu'il est ou faut-il en revoir les fondements? Les taux de sous-dénombrement apparaissent plus élevés du recensement de 1986, et ses potentielles répercussions sur les estimations démographiques, ont animé la discussion relative à la recherche d'une nouvelle méthode comme alternative à celle basée sur le recensement présentement en usage pour produire les estimations. Cette alternative ne s'appuierait plus nécessairement sur les données du plus récent recensement, mais utiliserait plutôt l'information pertinente disponible, y compris les chiffres administratifs, pour produire les estimations les «plus justes» possibles. En d'autres termes, les données du recensement continueraient d'entrer comme une part importante dans le processus d'estimation, mais pas au détriment d'autres éléments; pas plus d'ailleurs que le recensement le plus récent ne devrait avoir la prépondérance si, par exemple, le précédent était jugé plus fiable.

Après un examen attentif de la situation, Statistique Canada a décidé d'utiliser les données du recensement de 1986 (non corrigées des effets du sous-dénombrement) pour produire les estimations postcensitaires de 1986 et réviser les estimations intercensitaires de la période de 1981 à 1986. En d'autres termes, on a endossé les méthodes d'estimation existantes. Toutefois, il importe également qu'on accélère l'évaluation de la population à partir du recensement de 1991. Une telle stratégie devra prendre en compte les plans et les possibilités réelles d'amélioration et de développement dans les quatre domaines suivants:

- (1) couverture du recensement de 1991;
- (2) mesure du sous-dénombrement et du surdénombrement;
- (3) dossiers administratifs utilisés aux fins de l'élaboration des statistiques démographiques: amélioration des sources actuellement utilisées, à savoir les fichiers des allocations familiales et de l'impôt sur le revenu, et utilisation de nouvelles sources de données, comme le fichier de la sécurité de la vieillesse et les fichiers provinciaux de l'assurance-maladie;
- (4) estimations des migrations, particulièrement celles ayant trait à la migration interprovinciale, aux résidents canadiens de retour au pays après un séjour prolongé à l'étranger et à l'émigration.

Tableau 6

Estimations de l'émigration selon diverses méthodes, Canada, 1981 à 1986

Méthode	1981-86
Méthode résiduelle	
a) sans ajustement pour le sous-dénombrement	476,406
b) avec ajustement pour le sous-dénombrement	134,807
Fichier d'impôt de Revenu Canada	165,272
Méthode des allocations familiales (courante) (en utilisant le facteur f établi à partir du fichier d'impôt)	235,481
Méthode des allocations familiales (proposée) (en utilisant le facteur f établi à partir du fichier de l'immigration)	275,762
Contre-vérification des dossiers ¹	288,376

¹ Provisaires.
Source: Division de la démographie, Statistique Canada.

conversion, fondé sur les données de l'impôt sur le revenu. On utilise la même méthode pour le calcul des estimations provisoires de l'émigration et pour celui des estimations définitives, mais on utilise des données plus complètes dans le deuxième cas.

Le tableau 6 permet de comparer, pour la période intercensitaire de 1981 à 1986, les estimations de l'émigration fondées sur les fichiers d'allocations familiales et celles obtenues par les diverses autres méthodes mises en oeuvre. Il est frappant de constater que les estimations de l'émigration obtenues par la méthode résiduelle, que les chiffres du recensement aient été corrigés ou non, divergent des estimations plus plausibles dérivées des fichiers administratifs et de la contre-vérification des dossiers.

Bref, particulièrement depuis 1981, d'importantes innovations méthodologiques ont été apportées au système d'estimation de la migration interprovinciale et de l'émigration. Bien qu'on puisse supposer que ces innovations ont permis d'accroître la qualité générale des estimations, aucune preuve ne peut être produite à cet effet. Les données des allocations familiales et de l'impôt sur le revenu souffrent de diverses faiblesses inhérentes à tout système de données conçu pour des fins administratives plutôt que statistiques.

6. CONCLUSIONS ET QUESTIONS D'ACTUALITÉ

Le système d'estimation de la population de Statistique Canada repose sur deux ensembles fondamentaux de données : 1) les chiffres de population du recensement, et 2) les composantes de l'accroissement démographique, à savoir les naissances, les décès et les migrations. Les estimations postcensitaires sont obtenues en ajoutant à la population de base fournie par le recensement les composantes de la variation de la population au cours des années subséquentes. Elles sont rétrospectivement révisées lorsque les chiffres du recensement suivant deviennent disponibles. Ainsi, les données du recensement constituent à la fois la base des estimations postcensitaires et la norme qui sert rétrospectivement à leur validation. Le système a produit dans des délais relativement courts des estimations démographiques fiables et homogènes, et il s'est révélé d'une remarquable stabilité au fil des ans.

Cette stabilité peut être attribuée en majeure partie à la qualité élevée des recensements du Canada. Pour l'ensemble du Canada, le taux de sous-dénombrement mesuré par la contre-vérification des dossiers est resté presque inchangé, aux environs de 2%, pour trois recensements

Tableau 5
Migration interprovinciale nette pour la période de 1981 à 1986,
selon diverses sources

Unité géographique	Recensement de 1986 ¹	Allocations familiales	Impôt sur le revenu
Canada	0	0	0
Terre-Neuve	-16,550	-14,837	-15,051
Ile-du-Prince-Edouard	1,540	293	751
Nouvelle-Ecosse	6,275	5,204	6,895
Nouveau-Brunswick	-1,370	-2,239	-65
Québec	-63,295	-76,040	-81,254
Ontario	99,355	115,497	121,767
Manitoba	-1,555	-3,700	-2,634
Saskatchewan	-2,820	-668	-2,974
Alberta	-27,665	-34,073	-31,676
Colombie-Britannique	9,500	13,289	7,382
Yukon	-2,665	-2,381	-2,775
Territoires du Nord-Ouest	-755	-345	-366

¹ Population âgée de 5 ans et plus.
Source: Division de la démographie, Statistique Canada.

au fichier des allocations familiales, le fichier de l'impôt a l'avantage de reposer sur une base démographique beaucoup plus large : les déclarants et leurs dépendants représentent environ 90% de la population. Toutefois, les données sur la migration obtenues à partir des fichiers d'impôt sont sujettes à divers biais et erreurs. Les renseignements sur les personnes à la charge des déclarants doivent être déduits du montant de l'exemption personnelle totale. Diverses inférences sont nécessaires pour déterminer le statut migratoire des personnes à la charge des déclarants ainsi que des personnes qui ne remplissent pas de déclaration d'impôt sur le revenu et ne sont pas à la charge d'un déclarant, et donc, qui ne sont pas saisies par le système. Figurent notamment au nombre de ces personnes, les jeunes adultes et les personnes âgées plus susceptibles de négliger de remplir leur déclaration d'impôt sur le revenu ou de ne pas gagner le revenu minimal nécessaire pour remplir une déclaration. Si les données présentent réellement de tels biais liés à l'âge, ces derniers influent sur les estimations de la structure par âge qui, à leur tour, influent sur la valeur du facteur *f* utilisé aux fins des estimations provisoires de la migration interprovinciale reposant sur le fichier des allocations familiales.

Le tableau 5 fait état de la migration interprovinciale nette au cours de la période intermédiaire de 1981 à 1986 selon le fichier des allocations familiales, celui de l'impôt et la question du recensement concernant le lieu de résidence du répondant, cinq ans auparavant. Malgré certaines variations importantes de la valeur numérique des résultats obtenus, les trois sources de données permettent de brosser un tableau cohérent de l'ampleur de la migration interprovinciale nette par province, au cours de la période de cinq ans.

Les remarques relatives à la migration interprovinciale sont aussi valables pour l'émigration, c'est-à-dire sur les Canadiens qui s'établissent dans un autre pays. Avant 1981, l'émigration vers des pays autres que les États-Unis et le Royaume-Uni (dont les services d'immigration fournissaient les données pertinentes) devait être estimée de façon résiduelle à partir de deux recensements successifs et des composantes de l'accroissement intermédiaire de la population. Depuis 1981, l'estimation du nombre d'émigrants est fondée sur le fichier des allocations familiales et sur celui de l'impôt sur le revenu. La méthode utilisée est identique à celle décrite plus haut pour estimer la migration interprovinciale. L'émigration des enfants est déduite des données sur les allocations familiales, tandis que celle des adultes, et par conséquent l'émigration totale, est obtenue en appliquant à l'émigration des enfants un facteur de

Tableau 4

Répartition relative (en pourcentage) des immigrants par province selon le recensement de 1981 et les dossiers de l'immigration sur la province de destination prévue en 1980

Unité géographique	Dossiers de l'immigration	Recensement
Terre-Neuve	0.4	0.3
Ile-du-Prince-Edouard	0.1	0.1
Nouvelle-Ecosse	1.1	1.0
Nouveau-Brunswick	0.8	0.8
Québec	15.7	15.0
Ontario	43.5	42.7
Manitoba	5.4	5.4
Saskatchewan	2.5	2.6
Alberta	13.2	14.5
Colombie-Britannique + Yukon + Territoires du Nord-Ouest	17.2	17.6
Canada	100.0	100.0

Source: Division de la démographie, Statistique Canada.

les données des allocations familiales, ainsi que les estimations annuelles définitives, exploitant les données de l'impôt sur le revenu. Ces sources de données possèdent toutes deux leurs points forts et leurs limites.

Les principaux avantages des données sur les allocations familiales sont leur actualité et leur haut degré de précision. On peut disposer des données sur les changements d'adresse deux mois après le déménagement. La qualité des données du fichier repose sur deux facteurs. Le premier est le degré de complétude de la couverture de la population âgée de 18 ans ou moins, puisque tout enfant de ces âges à la charge de ses parents a droit de recevoir une allocation mensuelle. Le second est le stimulant financier incitant les bénéficiaires d'allocations familiales à déclarer tout changement d'adresse dès qu'il se produit. Toutefois, le fichier des allocations familiales ne contient aucun renseignement sur la migration des adultes. Cette dernière doit être estimée de façon indirecte en appliquant un facteur de conversion, f , égal au rapport du taux de migration des adultes à celui des enfants, tous deux obtenus à partir des fichiers d'impôt portant sur l'exercice financier le plus récent.

Compte tenu de l'importance du facteur f dans la formule d'estimation, il convient d'apporter quelques précisions à son sujet. Avant 1971, la valeur de f était calculée à l'aide de la déclaration, au plus récent recensement, du lieu de résidence cinq ans auparavant. Lorsqu'on a pu disposer, à partir des fichiers d'impôts, de données annuelles sur les migrants suivant l'âge, on a décidé de calculer f à l'aide de ces données qui ont l'avantage, par rapport à celles du recensement, de refléter un modèle migratoire plus récent.

Une autre innovation vaut la peine d'être soulignée à cet égard. Avant 1981, le facteur f était calculé seulement selon la province d'origine. Toutefois, lorsqu'on a disposé de données pertinentes tirées des fichiers d'impôt, il est apparu évident que le facteur variait aussi de façon significative selon la province de destination. On a donc décidé de calculer le facteur f selon la province d'origine et la province de destination.

Si on examine maintenant le fichier de l'impôt sur le revenu comme source de données pour l'estimation de la migration interprovinciale, les remarques suivantes s'imposent. Par rapport

5. QUE VALENT LES DONNÉES SUR LES COMPOSANTES DE L'ACCROISSEMENT DÉMOGRAPHIQUE?

Quelles conclusions pouvons-nous tirer de cette analyse de la variation intercensitaire de la population? Les deux séries d'estimations obtenues pour la période de 1976 à 1981 sont raisonnablement cohérentes. Les divergences sont peu marquées et se situent à l'intérieur de limites tolérables pour le Canada et pour la plupart des provinces. Il en va toutefois tout autrement pour les données relatives à la plus récente période intercensitaire, celle s'étendant de 1981 à 1986. Il semble qu'il y ait eu détérioration de la qualité des données et il reste à déterminer si ce sont les données des recensements ou celles des composantes de la croissance démographique qui sont touchées. Comme nous l'avons vu à la section précédente, deux méthodes différentes démontrent que le recensement de 1986 a enregistré un accroissement important du sous-dénombrement. Toutefois, la correction des effets du sous-dénombrement n'a pas toujours permis d'obtenir de meilleures estimations de la croissance démographique intercensitaire; de fait, dans certains cas, le contraire s'est produit. Dans la section suivante, nous allons examiner de plus près les données sur les composantes de l'accroissement démographique.

On trouve ci-après une brève évaluation de la qualité des données sur les naissances, les décès, l'immigration, l'émigration et la migration interprovinciale. Pour un compte rendu plus complet sur les données relatives aux composantes de l'accroissement démographique et sur les méthodes d'estimation des migrations, le lecteur est prié de se reporter à la publication de Statistique Canada «Méthodes d'estimation de la population, Canada».

On juge que l'enregistrement des naissances et des décès est complet au Canada. Compte tenu des règlements en vigueur concernant la déclaration des décès (obligation d'obtenir un permis d'inhumer) ainsi que des incitatifs matériels (allocations familiales) et des exigences légales favorisant l'enregistrement des naissances, le nombre des décès ou des naissances non déclarés ne peut qu'être très faible. Il est possible que surviennent certains retards dans l'enregistrement, mais le nombre de ces retards est peu élevé. Ainsi, pour la période de 1981 à 1985, 3,831 naissances, soit 0,02% de l'ensemble des naissances, et 2,528 décès, ou 0,03% de l'ensemble des décès, ont été déclarés au delà de la date limite, ce qui représente un bilan de 1,303 personnes non prises en compte par les estimations démographiques.

Dans la mesure où l'on parle des immigrants reçus, on considère que les statistiques d'immigration sont raisonnablement exactes. La répartition des immigrants par province est basée sur la province de destination prévue plutôt que sur la province de destination réelle. Il convient toutefois de souligner que cette répartition (voir tableau 4) correspond étroitement à la répartition établie à partir du recensement de 1986.

Comparativement aux trois composantes susmentionnées – naissances, décès et immigration – les migrations internes et l'émigration ne font l'objet d'aucun enregistrement direct. Les chiffres de l'équation (1) utilisée pour estimer la population au cours des années postcensitaires. En effet, les migrations internes et l'émigration ne font l'objet d'aucun enregistrement direct. Les chiffres correspondants doivent être estimés de façon indirecte à partir des fichiers administratifs – allocations familiales et impôt sur le revenu – renfermant des renseignements sur les changements de résidence. Ils méritent donc de faire l'objet d'un examen plus détaillé. Dans les paragraphes qui suivent, nous passerons en revue les importantes améliorations apportées au fil des ans aux méthodes et à la qualité des données et traiterons de certaines des imperfections inhérentes à ces estimations. (Pour un compte rendu plus complet, se reporter aux chapitres IV et V de Statistique Canada 1987).

Bien qu'on utilise les données sur les allocations familiales depuis 1956, l'innovation la plus importante apportée au système d'estimation de la migration interprovinciale a été le recours aux données de l'impôt sur le revenu en 1976. Depuis 1981, on élabore deux séries d'estimations des flux migratoires: les estimations provisoires, trimestrielles et annuelles, reposant sur

Rapport entre l'estimation fondée fur le recensement et celle fondée sur les composantes de l'accroissement démographique, par province, 1976 à 1981 et 1981 à 1986

Tableau 3

Rapport entre les estimations fondées sur les recensements et celles fondées sur les composantes de l'accroissement démographique x 100		Unité géographique			
1976-81	1981-86	Non corrigé des effets	Non corrigé du sous-dénombrement	Corrigé des effets du sous-dénombrement	Corrigé des effets du sous-dénombrement
		Non corrigé	Non corrigé	Non corrigé	Non corrigé

Canada	80.9	108.4	104.5	106.1
(Excluant les Territoires)				
Terre-Neuve	5.7	19.6	58.3	80.9
Ile-du-Prince-Edouard	76.7	101.6	109.8	135.7
Nouvelle-Ecosse	70.4	110.3	101.3	111.1
Nouveau-Brunswick	55.6	86.7	111.3	99.2
Québec	54.2	97.1	122.7	83.8
Ontario	88.1	115.2	92.0	103.2
Manitoba	89.2	117.3	35.6	29.0
Saskatchewan	79.4	110.3	112.0	105.5
Alberta	88.8	94.5	115.6	124.4
Colombie-Britannique	89.5	118.3	102.2	105.8

Nota: On ne peut établir cette comparaison pour la période de 1971 à 1976 et pour les périodes précédentes, puisque les estimations de l'émigration étaient alors produites par la méthode résiduelle à partir des résultats de deux recensements consécutifs et de l'accroissement résultant des autres composantes (naissances, décès et immigrants). Source: Division de la démographie, Statistique Canada.

interprovinciale (fichiers des allocations familiales et de l'impôt sur le revenu ainsi que la question du recensement portant sur la mobilité), ce déficit varierait entre 14,800 et 16,500 per-

sonnes (voir le tableau 5). On observe de semblables incohérences dans le cas du Québec. Pour la période de 1981 à 1986, l'estimation fondée sur les recensements, qui représente seulement 64% de celle fondée sur les composantes, impliquerait des pertes dues à l'émigration deux fois supérieures à celles estimées par Statistique Canada, soit de 160,000 plutôt que de 80,000 personnes. Or, encore une fois, selon les trois sources de données migratoires dont on dispose, le déficit net se chiffrait entre 63,000 et 81,000 personnes pour cette période de cinq ans. Toutefois, l'écart entre les deux estimations de la variation de la population disparaît presque entièrement lorsque les chiffres des recensements de 1981 et 1986 sont corrigés des effets du sous-dénombrement.

Enfin, le Nouveau-Brunswick tombe dans le même camp que le Québec et Terre-Neuve. Pour la période de 1981 à 1986, l'estimation fondée sur les recensements indique un déficit migratoire de 11,200 personnes, alors que la perte ne serait que de 2,200 personnes selon le fichier des allocations familiales. Qui plus est, les chiffres correspondants ne seraient que de 1,376 et 65 personnes respectivement selon la question du recensement sur la mobilité et selon le fichier de l'impôt sur le revenu. Une fois corrigées des effets du sous-dénombrement, les deux estimations des variations interconsistantes de la population pour le Nouveau-Brunswick se situent bien en deçà de la limite de tolérance.

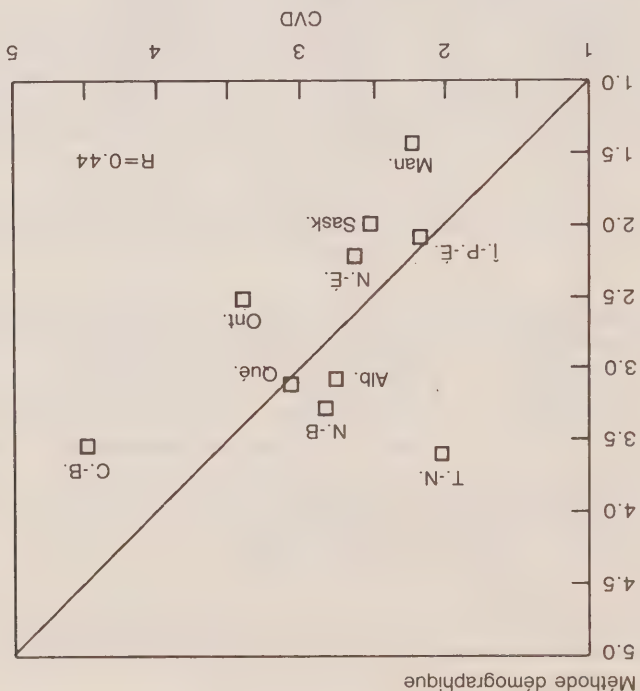


Figure 1. Relation entre les taux de sous-dénombrement estimés par la contre-vérification des dossiers et la méthode démographique, recensement de 1986

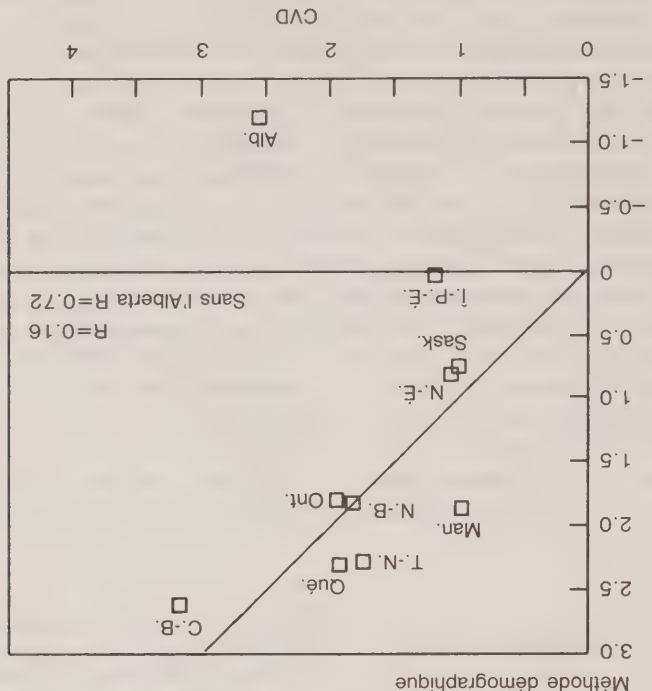


Figure 2. Relation entre les taux de sous-dénombrement estimés par la contre-vérification des dossiers et la méthode démographique, recensement de 1981

4. VARIATION INTERCENSITAIRE DE LA POPULATION CALCULÉE À PARTIR DES RECENSEMENTS ET DES COMPOSANTES DE L'ACCROISSEMENT: UN CONTRÔLE DE LA COHÉRENCE

Il reste à comparer entre eux deux ensembles indépendants d'estimations de la variation intercensitaire de la population: l'un fondé sur les composantes de l'accroissement démographique (naissances, décès et migrations), et l'autre sur les résultats de deux recensements consécutifs, non corrigés ou corrigés des effets du sous-dénombrement. Ces deux ensembles d'estimations seront respectivement nommés *estimations fondées sur les composantes* ($G_{t-s,t}$) et *estimations fondées sur les recensements* ($G_{t-s,t}^r$).

$$(9) \qquad G_{t-s,t} - B_{t-s,t} - D_{t-s,t} + I_{t-s,t} - E_{t-s,t} + N_{t-s,t}$$

$$(10) \qquad G_{t-s,t} = R_t - R_{t-s}$$

$$(11) \qquad G_{t-s,t}^r = (R_t + U_t) - (R_{t-s} + U_{t-s}).$$

La notation utilisée a été définie dans les formules précédentes. On peut considérer que deux estimations obtenues indépendamment sont raisonnablement fiables si elles sont similaires pour un point donné dans le temps. Comme on peut le voir au tableau 3, la différence entre les estimations fondées sur les recensements et celles fondées sur les composantes n'est que d'environ 5% pour la période de 1976 à 1981. Pour la période de 1981 à 1986, les deux ensembles divergent par une marge substantielle de 19% si elles ne sont pas corrigées des effets du sous-dénombrement, et par une marge de 8% si elles en sont corrigées. Il est plus délicat de comparer ces données à l'échelon provincial. Dans le cas des estimations fondées sur les composantes, c'est la fiabilité des estimations de la migration interveni-
ciale qui pose problème; dans celui des estimations fondées sur les recensements, c'est la variabilité des biais relatifs au sous-dénombrement et au surdénombrement, ainsi que les erreurs d'échantillonnage afférentes aux estimations du sous-dénombrement par la contre-vérification des dossiers. Dans certaines provinces, les seules erreurs d'échantillonnage pourraient com-
pter pour jusqu'à 15% des changements du rapport entre les deux estimations des variations intercensitaires de la population. Il est probable que l'explication de toute variation supérieure à 15% doive faire intervenir des erreurs ou des biais résultant de facteurs autres que l'échan-
tillonnage.

Par conséquent, en l'absence d'un critère plus fiable, nous avons fixé à $\pm 15\%$ la tolérance pour les divergences entre les deux estimations. Cette limite a au moins le mérite de nous permettre d'éliminer les cas trop douteux.

En ayant ces réserves en tête, examinons maintenant le tableau 3, qui compare par prov-
ince les estimations fondées sur les recensements et celles fondées sur les composantes pour les deux dernières périodes intercensitaires. Pour la période de 1976 à 1981, six provinces sur dix subissent ce test de tolérance quelque peu arbitraire avec succès; cette proportion passe à quatre provinces sur dix pour la période de 1981 à 1986. En général, les divergences observées sont plus marquées pour la période de 1981 à 1986 que pour celle de 1976 à 1981. À cet égard, Terre-Neuve, le Québec et le Nouveau-Brunswick sont particulièrement remarquables.

Dans le cas de Terre-Neuve, l'estimation des variations de la période de 1981 à 1986 fondée sur le recensement représente seulement 5% de celle fondée sur les composantes, et, même aussi faible ne saurait s'expliquer que par un déficit migratoire net d'environ 26,000 per-
sonnes pour cette période de cinq ans. Or, selon les trois sources de données sur la migration

Tableau 2
Taux de sous-dénombrement estimés par la méthode démographique
et par la contre-vérification des dossiers,
par province, 1981 et 1986

Unité géographique	Méthode démographique		Contre-vérification des dossiers ¹	
	1981 (%)	1986 (%)	1981 (%)	1986 (%)

Canada	2.82	2.01	(0.09)	3.21	(0.12)
(Excluant les	1.70	2.29	(0.95)	2.01	(0.32)
Terre-Neuve	0.05	2.10	(0.54)	2.16	(0.80)
Ile-du-Prince-Edouard	0.82	2.22	(0.34)	2.63	(0.38)
Nouvelle-Ecosse	1.83	3.28	(0.30)	2.83	(0.36)
Nouveau-Brunswick	2.31	3.13	(0.21)	3.06	(0.29)
Québec	1.81	2.53	(0.14)	3.40	(0.19)
Ontario	1.88	1.44	(0.35)	2.22	(0.40)
Manitoba	0.76	2.00	(0.37)	2.51	(0.36)
Saskatchewan	-1.18	3.09	(0.36)	2.75	(0.33)
Alberta	2.62	3.55	(0.33)	4.49	(0.39)
Colombie-Britannique					

¹ Les chiffres entre parenthèses indiquent l'écart-type.
Source: Division de la démographie, Statistique Canada.

Tout en gardant à l'esprit les remarques concernant les biais et les différences conceptuelles, voyons maintenant jusqu'à quel point les deux mesures de la couverture sont cohérentes à l'échelon provincial. À cette fin, nous utiliserons le critère de cohérence suivant: si deux mesures de la couverture sont identiques du point de vue conceptuel et empiriquement exactes, les points de corrélation des deux mesures dans l'espace doivent s'aligner le long de la bissectrice. Pour le recensement de 1981, si on excepte le cas de l'Alberta mentionné plus haut (et l'Ile-du-Prince-Edouard, fortement touchée par l'erreur d'échantillonnage), les points suivent de près la droite théorique. Les divergences observées sont peu importantes: dans la plupart des cas, elles n'ont aucune signification statistique compte tenu de l'écart-type des estimations obtenues par la contre-vérification des dossiers (voir le tableau 2).
Pour le recensement de 1986, les points de six provinces sur dix (Saskatchewan, Nouvelle-Ecosse, Ile-du-Prince-Edouard, Québec, Alberta et Nouveau-Brunswick) se rapprochent considérablement de la bissectrice et subissent donc avec succès le test de cohérence. Le point de Terre-Neuve s'éloigne considérablement de la bissectrice du côté gauche, ce qui laisse supposer une sous-évaluation du sous-dénombrement estimé par la contre-vérification des dossiers pour cette province. Par ailleurs, le Manitoba, l'Ontario et la Colombie-Britannique sont situés loin à droite de la bissectrice, ce qui suggère ou une surestimation du sous-dénombrement par la contre-vérification des dossiers ou une sous-évaluation du taux de couverture par la méthode démographique.
Encore une fois, il importe de souligner que l'analyse de l'exacitude de la couverture du recensement a été entravée par l'absence de données sur le surdénombrement. Néanmoins, il est juste de dire que, en dépit de ses limites, l'analyse indique clairement une détérioration de la couverture au recensement de 1986.

En outre, le taux de sous-dénombrement, n , calculé à l'aide de l'équation (6), accuse une légère distorsion à la baisse parce que la composante R'_t du dénominateur inclut un surdénombrement dont on ignore l'importance. À la lumière de ce qui précède, on voit que ces deux mesures de la qualité de couverture ne sont pas entièrement comparables.

Cependant, il y a également des différences conceptuelles. L'estimation obtenue par l'intermédiaire de la contre-vérification des dossiers constitue une mesure directe du sous-dénombrement, tandis que l'estimation démographique de l'erreur de couverture représente une entité plus complexe, difficile à définir sans équivoque. Elle n'est une mesure ni du sous-dénombrement, ni du sous-dénombrement net. Afin de mieux saisir la relation entre les deux estimations, l'équation (3) de la population attendue, P'_t , peut être exprimée comme en (7). On notera que la population recensée, R , est maintenant remplacée par ses deux composantes: les personnes correctement recensées, R' , et les doubles comptes, O .

$$P'_t = \left[(R'_{t-s} + O_{t-s}) + U_{t-s} \right] + G_{t-s,t} \quad (7)$$

L'équation (5), exprimant le taux de sous-dénombrement estimé par la méthode démographique, devient:

$$u'_t = \frac{[(R'_{t-s} + O_{t-s}) + U_{t-s} + G_{t-s,t}] - (R'_t + O_t)}{(R'_{t-s} + O_{t-s}) + U_{t-s} + G_{t-s,t}} \quad (8)$$

On voit, à la lumière de l'équation (8), que le biais introduit par le surdénombrement influe aussi bien sur le total de la population attendue que sur celui de la population recensée. En conséquence, l'estimation démographique du taux de sous-dénombrement reflète à la fois le sous-dénombrement comme tel et la différence entre le surdénombrement, O , au temps $t - 5$ (recensement de base) et au temps t (recensement terminal). Si on suppose exacts (a) le sous-dénombrement, aux temps $t - 5$, et t , du recensement estimé par la contre-vérification des dossiers, n , et, (b) l'accroissement démographique intercensitaire (bilan des composantes), $G_{t-s,t}$, alors le taux de couverture obtenu par la méthode démographique, u'_t , et celui du sous-dénombrement résultant de la contre-vérification des dossiers, u_t , varieront numériquement en fonction du surdénombrement aux recensements effectués en $t - 5$, et t , de sorte que si $O_t \geq O_{t-s}$ alors $u'_t \geq u_t$.

Ayant clarifié les particularités conceptuelles des deux mesures du sous-dénombrement, nous pouvons passer à l'examen du tableau 2, qui présente les deux types d'estimation du taux de couverture des recensements du Canada de 1981 et 1986. Tous deux indiquent un accroissement important de l'erreur de couverture dans le recensement de 1986. Toutefois, la méthode démographique produit dans l'ensemble des taux moins élevés que ceux obtenus par la contre-vérification des dossiers: ces taux sont respectivement de 2.82% et 3.21% pour 1986, et de 1.70% et 2.01% pour 1981. On peut en conclure que le surdénombrement était plus marqué en 1981 qu'en 1976 et plus élevé en 1986 qu'en 1981, pour autant que les hypothèses qui sous-tendent les relations soient correctes. Malheureusement, nous ne disposons d'aucune donnée pouvant confirmer ou infirmer ces hypothèses.

On trouve aux figures 1(a) et 1(b) une représentation des estimations de l'erreur de couverture par province (selon la méthode démographique et selon la contre-vérification des dossiers) figurant au tableau 2. L'application des différences observées à l'échelon provincial est plus incertaine encore parce qu'on peut s'attendre à ce que les erreurs et les biais mentionnés ci-dessus soient plus sensibles à cet échelon qu'à l'échelon national. Ceci est particulièrement vrai pour l'erreur d'échantillonnage relative aux estimations dérivées de la contre-vérification des dossiers et pour les biais associés à la mesure de la migration interprovinciale qui compromettent l'exactitude de la mesure des variations intercensitaires de population dans le cas de la méthode démographique.

3. OBTENTION DU TAUX DE SOUS-DÉNOMBREMENT PAR LA MÉTHODE DÉMOGRAPHIQUE

En corrigeant le recensement de base des effets du sous-dénombrement tel qu'estimé par la contre-vérification des dossiers et en lui ajoutant l'accroissement démographique (excédent des naissances sur les décès et migrations) de la période postcensitaire, on obtient, conformément à l'équation 3, le chiffre de la population à la date du recensement suivant. Nous parlerons alors de population *attendue*, par opposition à la population *estimée* et à la population *recensée*.

$$\text{ou: } P'_t = \left[R_{t-5} + U_{t-5} \right] + G_{t-5,t}, \quad (3)$$

P'_t = population attendue au temps t ;

R_{t-5} = population recensée au temps $t-5$;

U_{t-5} = nombre de personnes omises dans le recensement $t-5$, tel qu'obtenu par la contre-vérification des dossiers;

G_{t-5} = estimation de l'accroissement intercensitaire de la population pendant la période $t-5, t$ (naissances, décès et migrations de l'équation (1)).

La différence, U'_t , entre la *population attendue*, P'_t , et la *population recensée*, R_t , (équation 4) peut être considérée comme une erreur de couverture. Nous appellerons cette différence l'*estimation démographique* de l'erreur de couverture.

$$U'_t = P'_t - R_t. \quad (4)$$

Le taux de l'erreur de couverture, u'_t , est simplement le rapport de l'estimation démographique de l'erreur de couverture, U'_t , à la population attendue, P'_t :

$$u'_t = \frac{P'_t}{P'_t - R_t} = \frac{P'_t}{U'_t}. \quad (5)$$

Par comparaison, le taux de sous-dénombrement résultant de la contre-vérification des dossiers s'obtient comme suit:

$$u_t = \frac{R_t}{U_t + R_t}. \quad (6)$$

Quels rapprochements peut-on faire entre l'estimation démographique de l'erreur de couverture et celle du sous-dénombrement par l'intermédiaire de la contre-vérification des dossiers? Tout d'abord, il importe de souligner que toutes deux sont sujettes à des erreurs et à des biais. La première subit les effets : (a) de l'absence d'une estimation du surdénombrement; (b) de l'erreur attachée au facteur U utilisé pour corriger les recensements de base et terminal des effets du sous-dénombrement; et (c) des biais associés à l'estimation de la croissance démographique intercensitaire, $G_{t-5,t}$, particulièrement celle de la composante migratoire. L'estimation du sous-dénombrement par la contre-vérification des dossiers est touchée par : (a) l'erreur d'échantillonnage; et (b) divers biais liés au dépistage des personnes, à l'appariement des dossiers, etc.

que le signe variait en 1971, 1976 et 1981. En outre, on peut remarquer que la valeur de l'erreur enregistrée en 1986 est plus élevée dans la plupart des provinces que dans les trois provinces précédentes. Les erreurs en fin de période les plus importantes caractérisaient les provinces maritimes et le Québec, tandis que l'Ontario et les provinces de l'Ouest, à l'exception de la Saskatchewan, affichaient les erreurs les plus faibles.

Il convient de formuler d'autres remarques en ce qui a trait à l'erreur observée pour l'Alberta en 1981. Cette année-là, la province a affiché une erreur en fin de période exceptionnellement élevée: les estimations se sont révélées de 2,41 % inférieures aux chiffres du recensement, ce qui représente une différence de 53,886 individus. Ces résultats peuvent s'expliquer de deux façons. D'une part, il est possible que le recensement de 1981 ait donné lieu à un (sur)dénombrement important dans cette province. Attirés par l'essor économique engendré par l'exploitation des gisements pétroliers de la province, un grand nombre de personnes originaires d'autres provinces se sont rendues en Alberta à la recherche d'un emploi et certaines d'entre elles ont pu être recensées par erreur comme des résidents habituels de cette province. Pourtant, l'Alberta a affiché un taux de sous-dénombrement (2,54 %) supérieur à la moyenne en 1981, ce qui ne fait qu'ajouter à l'énigme. D'autre part, il peut arriver que le flux de personnes ayant migré en Alberta au cours de cette période de prospérité économique et d'explosion démographique n'ait pas été pleinement saisi par les fichiers des allocations familiales et de l'impôt sur le revenu sur lesquels se fondent les estimations de la migration interprovinciale. En d'autres termes, il est possible que la notable sous-estimation de la population par rapport au recensement de 1981 ait été le résultat d'une sous-évaluation de la migration vers l'Alberta dans l'estimation postcensitaire.

La preuve étant faite que l'écart entre les estimations et le recensement s'est élargi de façon considérable en 1986, il reste à déterminer si ce phénomène est attribuable à une détérioration: (a) de la couverture du recensement ou (b) de la qualité des données sur les composantes de l'accroissement de la population au cours de la dernière période intercensitaire.

Tableau 1

Erreur en fin de période: Canada, provinces et territoires juin 1971, 1976, 1981 et 1986

Unité géographique	Erreur (pour-cent) ¹			
	1971	1976	1981	1986
Canada	0.51	0.58	-0.25	0.95
Terre-Neuve	0.32	-0.19	1.25	2.02
Ile-du-Prince-Edouard	-0.76	1.58	-0.31	1.06
Nouvelle-Ecosse	-2.45	0.93	-0.03	1.28
Nouveau-Brunswick	-0.44	1.51	-0.28	1.57
Québec	0.08	0.10	-0.58	1.34
Ontario	1.41	1.07	0.37	0.73
Manitoba	-0.01	1.21	0.83	0.57
Saskatchewan	0.21	0.91	-0.52	1.06
Alberta	0.31	-0.09	-2.41	0.81
Colombie-Britannique	0.47	0.07	-0.22	0.58
Yukon	-6.63	-2.34	-2.11	-4.66
Territoires du Nord-Ouest	3.14	-0.92	-5.60	-1.32
¹ Estimation - Recensement x 100				
Recensement				

Source: Division de la démographie, Statistique Canada.

des décès et des émigrants. Pour estimer la population par province, on ajoute une composante, la migration interprovinciale nette. Ces calculs sont répétés chaque année pendant la période de cinq ans menant au prochain recensement. La méthode d'estimation actuelle prévoit une révision rétrospective des estimations postcensitaires afin de les aligner sur les chiffres du recensement qui les suit (Statistique Canada 1987). La différence, calculée selon l'équation 2, entre les estimations postcensitaires et le recensement est appelée ('erreur en fin de période) (EF).

$$(1) \quad P_t = R_{t-5} + \left[B_{t-5,t} - D_{t-5,t} + I_{t-5,t} - E_{t-5,t} + N_{t-5,t} \right]$$

$$(2) \quad EF(\%) = \frac{P_t - R_t}{R_t} \times 100,$$

où:

P_t = estimation de la population au temps t ;
 R = chiffres du recensement au temps t ou $t-5$ selon le cas;
 B = nombre de naissances;
 D = nombre de décès;
 I = nombre d'immigrants;
 E = estimation du nombre d'émigrants;
 N = estimation du solde migratoire interprovincial;
 $t-5, t$ indique la période de cinq ans sur laquelle portent les calculs.

Le tableau 1 présente l'erreur en fin de période afférente aux quatre derniers recensements pour le Canada et les provinces. En général, la correspondance entre les chiffres du recensement et les estimations est très bonne, même au niveau provincial. Cette situation est d'autant plus remarquable que, en l'absence d'enregistrement direct de ces mouvements, l'émigration et la migration interprovinciale doivent toutes deux être estimées à partir de données administratives (fichiers des allocations familiales et de l'impôt sur le revenu). Malgré la correspondance étroite entre les deux séries de données, l'erreur en fin de période observée présente deux caractéristiques frappantes. La première a trait au bond considérable qu'a enregistré l'erreur pour passer, en 1986, à près de un pour cent, pourcentage relativement important compte tenu des erreurs observées lors des recensements antérieurs. L'erreur observée à l'occasion des recensements de 1971 et de 1976 était légèrement supérieure à un demi de un pour cent et elle se chiffrait seulement à un quart de un pour cent en 1981. La seconde caractéristique a trait au signe de l'erreur en fin de période enregistrée en 1981. En effet, tandis que les estimations se sont révélées supérieures aux chiffres du recensement à l'occasion des trois autres recensements, c'est l'inverse qui s'est produit en 1981. La presque totalité de cet écart négatif était attribuable à l'estimation produite pour l'Alberta.

Si on examine les chiffres relatifs aux provinces, on constate qu'on a systématiquement enregistré une erreur en fin de période de signe positif pour toutes les provinces en 1986, tandis

de sources indépendantes ou encore avec des estimations calculées par des méthodes statistiques ou démographiques. À la suite des travaux pionniers de Ansley Coale (1955), le US Bureau of the Census a utilisé les techniques de couverture du recensement (se reporter au plus récent rapport de Fay *et al.* 1988). Certaines tentatives ont aussi déjà été faites dans ce sens au Canada (Lapierre 1970). Comme nous le verrons plus loin, l'essence de la méthode démographique consiste à tenir compte des relations formelles qui existent entre le chiffre de population et les composantes de l'accroissement démographique, à savoir les naissances, les décès et les migrations.

La couverture du recensement de 1986 a déjà fait l'objet d'une évaluation au moyen de la contre-vérification des dossiers et un rapport en rend compte (Carter 1988; Statistique Canada 1988). Qu'il suffise de souligner que les estimations du sous-dénombrement fondées sur cette évaluation sont sujettes à des erreurs d'échantillonnage, qui peuvent se révéler assez importantes pour les provinces peu peuplées, et à des biais dont on ignore l'ampleur (difficultés à retracer les personnes ou à apparier les dossiers). De plus, la contre-vérification des dossiers a été conçue avant tout pour mesurer le sous-dénombrement. On a tenté, à titre expérimental, d'utiliser cette méthode pour mesurer le surdénombrement, mais nous ne disposons pas des résultats de cet essai au moment d'écrire ces lignes. Pour ces raisons et d'autres semblables, la mise en oeuvre d'une mesure différente de l'exactitude des chiffres du recensement revêt d'autant plus d'importance.

Le présent document porte sur l'évaluation, au moyen de l'analyse démographique, des résultats des trois derniers recensements, et plus particulièrement du recensement de 1986. Cette évaluation s'effectue en trois étapes. Premièrement, on compare les chiffres du recensement et des estimations. Deuxièmement, des techniques démographiques sont mises en oeuvre pour produire d'autres estimations du sous-dénombrement au recensement qui sont à leur tour comparées avec celles résultant de la contre-vérification des dossiers. Enfin, la troisième et dernière étape consiste à faire porter l'évaluation sur la variation intercensitaire de la population plutôt que sur les effectifs globaux. Deux ensembles indépendants d'estimations de cette variation sont produits. Le premier est fondé sur l'utilisation des deux recensements encadrants (dits (de base) et (terminal)), tandis que le deuxième résulte directement des données sur les naissances, les décès et les migrations.

Avant d'entreprendre l'évaluation comme telle, une mise en garde s'impose. Bien qu'ils soient d'une qualité acceptable pour la plupart des utilisations qu'on en fait, ni les estimations de population ni les chiffres du recensement ne sont parfaits. De fait, il n'existe pas d'ensemble de données jugé assez parfait pour servir de référence aux fins de la validation d'autres données. Le fait est que les données statistiques sont toujours imparfaites à divers degrés. Dans l'état actuel des choses, il se peut qu'il soit impossible d'obtenir le degré d'affinement et de précision qu'exigent certaines utilisations – comme l'affectation des ressources et le transfert de recettes mentionnés plus haut. Toutefois, nous espérons que la présente évaluation, fondée sur la combinaison d'outils statistiques variés, quoique imparfaits, nous permettra de saisir la direction et l'importance de l'erreur et des biais entachant le recensement et diverses composantes des estimations de population. Souhaitons que la présente initiative nous permette d'améliorer les méthodes mises en oeuvre dans le cadre de la préparation du recensement de 1991 et serve aux fins du calcul des estimations démographiques postcensitaires subséquentes.

2. POPULATION RECENSÉE VERSUS ESTIMATIONS POSTCENSITAIRES: ERREUR EN FIN DE PÉRIODE

Les estimations démographiques postcensitaires sont obtenues, conformément à l'équation 1, à l'aide de la méthode dite des composantes qui consiste à ajouter le nombre des naissances et des immigrants au chiffre de population du recensement de base, et à en soustraire le nombre

Une approche démographique à l'évaluation du recensement de 1986 et des estimations de population pour le Canada

ANATOLE ROMANIUC¹

RÉSUMÉ

Un accroissement significatif dans le sous-dénombrement du recensement de 1986 est révélé tant par la contre-vérification des dossiers que par la méthode démographique présentée dans ce document. Une attention particulière est portée à l'évaluation des différentes composantes de l'accroissement de la population, spécialement à la migration interprovinciale. Le texte conclut par un survol de deux méthodes différentes pour générer les estimations postcensitaires: celle couramment utilisée, basée sur le recensement, et un modèle flexible utilisant toutes les données pertinentes en plus du recensement.

MOTS CLÉS: Sous-dénombrement du recensement; estimations de la population; méthode des composantes démographiques.

1. INTRODUCTION

L'exactitude des résultats du recensement, et des estimations démographiques postcensitaires fondées sur ces derniers, constitue en soi une question importante. L'utilisation des chiffres de population dans la formule de calcul des transferts de recettes entre les divers paliers de gouvernement confère à cette question un caractère à la fois crucial et particulièrement délicat du point de vue politique (Fellegi 1980; Romaniuc et Raby 1980). La vigueur des débats au Canada et aux États-Unis sur l'opportunité de corriger le recensement pour le sous-dénombrement de la population, ainsi que les quelques poursuites judiciaires engagées aux États-Unis à ce sujet, témoignent de la portée politique et de la complexité technique de la question.

Pourtant, malgré les nombreux écrits en la matière et l'abondance des arguments invoqués tant en faveur de l'ajustement des données que contre, le débat reste ouvert (Keyfitz 1979 et 1980; Kish 1980; Freedman et Navidi 1986; Stoto 1987). En fin de compte, Statistique Canada (tout comme le US Department of Commerce) a décidé de ne pas corriger les chiffres des effets de sous-dénombrement, tout en réitérant son engagement de longue date à l'égard de la politique d'évaluation de la qualité des données (Wilk 1981). La diffusion des résultats de cette évaluation et de la méthodologie qui la sous-tend permet aux utilisateurs d'apporter les corrections qui répondent à leurs besoins propres, en toute connaissance des forces et des lacunes des chiffres du recensement et des estimations. Le présent mémoire a été rédigé dans l'esprit de cette politique d'évaluation de la qualité.

Fondamentalement, il existe deux approches possibles pour mesurer l'exactitude des chiffres du recensement. La première approche, qu'on peut qualifier de (micro-évaluation), implique la vérification d'un certain nombre de dossiers individuels, apparus cas par cas, afin de repérer les personnes qui ont été oubliées, recensées plus d'une fois, ou encore recensées bien que, par définition, elles n'auraient pas dû faire partie de l'univers du recensement. Le Programme postcensitaire du US Bureau of the Census et la contre-vérification des dossiers effectuée par Statistique Canada appartiennent tous deux à ce genre d'évaluation.

La deuxième approche, qu'on peut qualifier de (macro-évaluation), consiste à analyser les données agrégées: ainsi on peut comparer les chiffres du recensement avec les chiffres obtenus

¹ Anatole Romaniuc, Directeur, Division de la démographie, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

- PLAUT, G. (1985). La reconnaissance au Canada du statut de réfugié: propositions relatives à l'élaboration d'un nouveau processus. Comité permanent du travail, de la main-d'oeuvre et de l'immigration. STATISTIQUE CANADA (1976 à 1981). *Mortalité, liste sommaire des causes - La statistique de l'état civil*, Statistique Canada, vol. III, n° 84-206 au catalogue.
- STATISTIQUE CANADA (1979). *Statut de mobilité et caractéristiques générales de la population*, Statistique Canada, n° 92-834 au catalogue.
- STATISTIQUE CANADA (1982). *Chiffres de population, 1976 et 1981, circonscriptions électorales fédérales*, Statistique Canada, n° 99-908 au catalogue.
- STATISTIQUE CANADA (1983a). *Mobilité*. Statistique Canada, n° 92-907 au catalogue.
- STATISTIQUE CANADA (1983b). Contre-vérification des dossiers, Totalisations: Projet de la qualité des données. Recensement de la population et du logement de 1981 (non publié).
- STATISTIQUE CANADA (1984). Bande de contre-vérification des dossiers à grande diffusion - Guide de l'utilisateur - Recensement de la population et du logement de 1981, Statistique Canada.
- STATISTIQUE CANADA (1987). *Méthodes d'estimation de la population*. Statistique Canada, n° 91-528F au catalogue.
- STATISTIQUE CANADA (1988). Taux de sous-dénombrement provenant de la contre-vérification des dossiers de 1986. Bulletin d'information à l'intention des utilisateurs, Statistique Canada, n° 2.

De façon similaire,

$$e(M_{76}) = M_{76}[e(M_{71})] - M_{76}(O_{nf76}) + M_{c/r76} + M_{n76}.$$

L'erreur et une partie de la différence $C - C$ peut être transmise d'une CVD à l'autre par la base des « personnes non recensées » et par les cas de surdénombrement dans la base des personnes recensées. Cette erreur peut expliquer une partie importante de la différence $C_{81} - C_{81}$. L'effet sur $C - C$ peut être beaucoup plus considérable que sur M .

Le taux de l'erreur de couverture nette au recensement de 1981, pour les dix provinces, serait égal à:

$$M_{81} + M_{n81} - O_{81}; \frac{C_{81} + M_{81} + M_{n81} - O_{81}}{M_{81} + M_{n81}};$$

et le taux de sous-dénombrement serait:

$$M_{81} + M_{n81} - O_{81}. \frac{C_{81} + M_{81} + M_{n81} - O_{81}}{M_{81} + M_{n81}}.$$

L'estimateur utilisé dans la CVD est:

$$\frac{M_{81}}{C_{81} + M_{81}}.$$

Une valeur, même petite, de $M_{n81} - O_{81}$ pourrait engendrer un biais significatif dans les résultats de la CVD si ces résultats étaient utilisés comme estimation de l'erreur de couverture nette. Une valeur, même relativement petite, de $e(M_{81})$ pourrait créer un biais significatif dans les estimations du sous-dénombrement de la CVD: deux éléments potentiels de biais seraient transmis par la CVD précédente; un élément de biais viendrait de toute erreur de classement dans la CVD; et un autre élément de biais serait associé aux « personnes non recensées » ne figurant dans aucune des bases de la CVD. Bien entendu, ces éléments peuvent se faire contre poids dans une certaine mesure.

Un autre estimateur qui pourrait être retenu consisterait à utiliser C_{81} au lieu de C_{81} au dénominateur. Dans certaines conditions plausibles, l'utilisation de C_{81} produirait fort probablement des estimations moins biaisées au niveau national. Toutefois, ces conditions, qui touchent aux tailles relatives de $C_{81} - C_{81}$, O_{81} et M_{n81} ne tiendraient pas pour les provinces ni pour les estimations pour lesquelles le chiffre du recensement est inférieur à l'estimation de ce chiffre au moyen de la CVD.

BIBLIOGRAPHIE

BURGESS, R. D. (1986). Major issues and implications of tracing survey respondents, International Symposium on Panel Survey, Washington D.C.

GOSSSELIN, J.-F. (1976). The methodology of the 1971 Reverse Record Check, *Techniques d'enquête*, 2, 180-193, résumé en français.

Equation 3:

L'estimateur C_{81} devrait être = $C_{81} - C_{81}[e(M_{76})] - C_{c/re81} - R_{81} - T_{76/81} - S_{76/81} + M_{n81} - O_{81} + C(O_{nf81})$,

où

C_{81} = nombre de personnes dénombrées dans une des dix provinces au recensement de 1981,

$C_{81}[e(M_{76})]$ = partie de $e(M_{76})$ qui ne figure pas ou qui est surreprésentée (-) en C_{81} ,

$C_{c/re81}$ = sous ou surestimation (-) des «personnes recensées» dans un base due à une erreur de classification, de réponse ou d'échantillonnage et aux cas «non

dépistés» dans la CVD de 1981,

R_{81} = nombre de personnes à l'étranger au moment du recensement de 1976 et qui étaient au Canada au recensement de 1981,

$T_{76/81}$ = nombre de personnes qui ont migré d'un des territoires vers une des provinces entre 1976 et 1981,

$S_{76/81}$ = nombre net de personnes entrées dans les dix provinces, au jour du recensement, et qui ne figuraient pas dans une des bases de la CVD et dont on n'a

pas tenu compte ci-dessus (immigrants illégaux),

M_{n81} = nombre de personnes qui ne figuraient pas dans une des bases de la CVD et n'ont pas été recensées dans l'une ou l'autre province en 1981,

O_{81} = surdénumérement dans les dix provinces au recensement de 1981,

$C(O_{nf81})$ = nombre estimé de cas de surdénumérement dans les bases «Recensement», «Naissances» et «Immigrants» qui n'ont pas été décelés à la CVD de 1981

et dont on n'a pas tenu compte en C_{81} .

Ainsi,

$$C_{81} - C_{81} = -C_{81}[e(M_{76})] - C_{c/re81} - S_{76/81} + M_{n81} - O_{81} + C(O_{nf81}) - R_{76/81} - T_{76/81}$$

en supposant qu'il n'y a pas d'erreur en O_{nf81} .

Equation 4:

$$M_{81} - M_{81} = M_{81}[e(M_{76})] - M_{81}(O_{nf81}) + M_{c/re81} + M_{n81} = e(M_{81}).$$

où

$$M_{c/re81}$$

= sous ou (-) surestimation du nombre de personnes non recensées dans une base de la CVD due à une erreur de classement, de réponse ou d'échantillonnage et aux cas «non dépistés» à la CVD de 1981,

$M_{81}[e(M_{76})]$ = composante de $e(M_{76})$, représentée en M_{81} ,

$M_{81}(C_{nf81})$ = nombre estimé de cas de surdénumérement dans les bases «Recensement», «Naissances» et «Immigrants» qui n'ont pas été décelés à la CVD de 1981

et ne sont pas représentés en M_{81} .

Nota: Chaque terme de l'équation 2 contient un élément d'erreur associé à la classification,

à la réponse, à l'échantillonnage et aux cas «non dépistés», par exemple $C_{c/re81}$ et $M_{c/re81}$. La somme de ces erreurs est nulle. Dans les équations ci-dessus, ces éléments d'erreur ne comprennent pas l'erreur engendré par le surdénumérement et les cas de surdénumérement attribuables aux cas «non dépistés» (c'est-à-dire les personnes qui n'existent pas comptées comme ayant été dénombrées au recensement précédent). L'effet

du surdénumérement est inclus, par exemple, dans $C(O_{nf81})$ et $M(O_{nf81})$.

Annexe 2

Equations utilisées aux fins de la CVD et estimateur

Les estimations de la CVD de 1981 ont été évaluées à partir de quatre équations. La première définit simplement les bases de sondage ou la population de la CVD. La deuxième redéfinit l'échantillon en termes des résultats ou des estimations obtenues. La troisième définit la population dénombrée au recensement d'après l'estimation par CVD du nombre de personnes comptées. La quatrième définit les composantes de l'erreur dans l'estimation du nombre de «personnes non recensées».

Equation 1:

Taille de la population de la CVD = $C_{76} + M_{76} - e(M_{76}) + I_{76/81} + B_{76/81}$, où:

C_{76} = nombre de personnes dénombrées dans une des dix provinces au recensement de 1976,
 M_{76} = nombre de personnes non dénombrées dans une des dix provinces au recensement de 1976,
 $e(M_{76})$ = erreur (sous ou surestimation (-) des personnes) associée à M_{76} , l'échantillon de la «base des personnes non recensées»; c'est-à-dire $M_{76} = M_{76} + e(M_{76})$,
 $I_{76/81}$ = nombre d'immigrants admis dans une des dix provinces entre 1976 et 1981,
 $B_{76/81}$ = nombre de naissances déclarées dans une des dix provinces entre 1976 et 1981.

Equation 2:

Estimations de la CVD = $\hat{C}_{81} + \hat{C}_{f781} + \hat{M}_{81} + \hat{M}_{f781} + \hat{L}_{76/81} + \hat{A}_{81} + \hat{D}_{76/81} + \hat{O}_{f81}$ où

\hat{C}_{81} = nombre estimé de personnes comprises dans une des bases de la CVD et qui ont été dénombrées dans une des dix provinces au recensement de 1981,
 \hat{C}_{f781} = nombre estimé de personnes comprises dans une des bases de la CVD et qui ont été dénombrées dans l'un ou l'autre territoire au recensement de 1981,
 \hat{M}_{81} = nombre estimé de personnes comprises dans une des bases de la CVD et non dénombrées dans une des dix provinces au recensement de 1981,
 \hat{M}_{f781} = nombre estimé de personnes comprises dans une des bases de la CVD et non dénombrées dans l'un ou l'autre territoire au recensement de 1981,
 $\hat{L}_{76/81}$ = nombre estimé de personnes comprises dans une des bases de la CVD et qui ont émigré entre 1976 et 1981,
 \hat{A}_{81} = nombre estimé de personnes comprises dans une des bases de la CVD et qui étaient à l'étranger et n'avaient pas de «résidence habituelle» au Canada au moment du recensement de 1981,
 $\hat{D}_{76/81}$ = nombre estimé de personnes comprises dans une des bases de la CVD et qui sont décédées entre 1976 et 1981,
 \hat{O}_{f81} = nombre estimé de cas de surdénombrement décelés dans les bases «Recensement», «Immigrants» et «Naissances» de la CVD de 1981.

Annexe 1

Autres résultats de la CVD

Les résultats de la CVD de 1986 ont été publiés par Statistique Canada en 1988. Dans l'extrait ci-dessous, on trouvera les taux de sous-dénombrement, selon certaines caractéristiques, des recensements de 1981 et de 1986. On continue en ce moment d'analyser les estimations du sous-dénombrement (selon la province, le groupe d'âge, le sexe, l'état matrimonial, la langue maternelle et d'autres caractéristiques).

CVD de 1981 et 1986: Taux de sous-dénombrement selon certaines caractéristiques démographiques, Canada (10 provinces)

Caractéristique	Taux estimé de 1981		Taux estimé de 1986	
	Taux	E.T.	Taux	E.T.
%				
Sexe				
Hommes	2.37	0.13	3.91	0.16
Femmes	1.65	0.12	2.87	0.16
Groupe d'âge				
0-4	1.21	0.22	2.28	0.48
5-14	1.23	0.21	2.12	0.26
15-19	2.96	0.52	3.89	0.60
20-24	5.51	0.29	9.06	0.45
25-34	2.31	0.28	4.76	0.32
35-44	2.20	0.26	2.40	0.32
45-54	0.81	0.23	1.77	0.28
55-64	0.91	0.29	2.09	0.31
État matrimonial				
Marié(e)/séparé(e)	1.22	0.11	1.89	0.15
Divorcé(e)	5.10	1.03	7.07	1.07
Veuf (veuve)	0.64	0.39	2.68	0.51
Célibataire/jamais marié(e)	2.86	0.16	4.91	0.21
Langue maternelle				
Anglais	1.86	0.11	3.12	0.13
Français	1.80	0.20	3.10	0.33
Autre	3.08	0.26	-	-
Groupe de taille de population				
Régions urbaines	2.08	0.11	3.28	0.13
500,000 et plus	2.29	0.17	3.58	0.15
100,000 à 499,999	1.86	0.31	2.94	0.33
Moins de 100,000	1.80	0.23	-	-
Régions rurales				
	1.79	0.21	3.73	0.29

géographiques dans la qualité de la couverture, révélées par la CVD, posent des difficultés. La variation dans la distribution des chiffres du recensement a été examinée afin de déterminer si elle est liée à la variation apparente du sous-dénombrement entre les provinces. Jusqu'à maintenant, ces examens n'ont pas permis d'établir des modèles adéquats ni de fournir des explications satisfaisantes.

L'incapacité de modéliser le sous-dénombrement ou d'en expliquer la variation interprovinciale peut être liée aux facteurs suivants: (i) biais (ou erreur d'échantillonnage) dans les estimations de la CVD; (ii) trop faible corrélation entre le sous-dénombrement et les caractéristiques démographiques des personnes, des ménages ou des familles; (iii) sous-dénombrement associé à un ensemble peut-être complexe de caractéristiques de la population recensée et d'autres facteurs; et/ou (iv) multitude de causes de sous-dénombrement à considérer séparément, par exemple les causes du sous-dénombrement des personnes d'une part et celles des ménages d'autre part.

4. CONCLUSION

On estime que la CVD est la meilleure méthode mise au point jusqu'à maintenant pour estimer le sous-dénombrement dans le recensement du Canada. Les estimations de la CVD consistent une mesure de base à partir de laquelle on peut évaluer la qualité des chiffres du recensement.

Toutefois, la CVD, telle qu'on l'applique en ce moment pour évaluer la couverture du recensement, comporte un certain nombre de faiblesses conceptuelles, théoriques et pratiques. Les listes ou bases dont est tiré l'échantillon couvrent la quasi-totalité de l'univers des personnes à dénombrer au recensement, mais elles ne sont pas complètes. Certaines régions géographiques et certaines tranches de la population sont exclues. La taille de l'échantillon est limitée – bien que ce ne soit pas forcément à sa valeur actuelle – par les contraintes du dépistage et de l'appariement et par les exigences d'exactitude dans les opérations. Les personnes sélectionnées qu'on n'arrive pas à dépister engendrent un biais dans les estimations. La proportion de personnes «non dépistées» par rapport aux personnes non recensées compromet la qualité des estimations de la CVD tout comme les différences entre le nombre estimatif de personnes recensées établi avec cette méthode et les chiffres officiels du recensement.

Dans certain cas, l'effet des sources d'erreur ou des faiblesses de la CVD pourraient être analysées de manière plus approfondie. D'autres méthodes (nouvelles ou modifiées) qui offrent des chances raisonnables d'améliorer la qualité et l'applicabilité des estimations de la CVD pourraient également être utilisées. En principe, on est en mesure de mettre au point de nouvelles méthodes, mais leur coût et leur efficacité varieraient. Et il resterait à savoir si de tels changements auraient d'autres effets que celui d'améliorer les estimations de la CVD comme indicateurs généraux de la qualité de chiffres du recensement.

REMERCIEMENTS

L'auteur remercie Gordon Brackstone, Geoff Hole, Judy Clarke et tout le personnel de Division des méthodes d'enquêtes sociales pour leur aide et leur commentaires dans la rédaction de cet article. Il tient également à remercier les arbitres et rédacteurs pour leurs observations.

en 1976 et en 1981, ne peut pas être attribué à un seul facteur. Si c'était le cas, il faudrait que la différence dans le niveau du surdéveloppement entre ces recensements ait été supérieure à la différence entre les estimations. Les analyses démographiques qui ont été faites pour les trois recensements (Statistique Canada 1987) ne confirment pas cette hypothèse.

Sans infirmer les estimations produites dans le cadre des analyses démographiques, on pourrait expliquer en partie les différences par l'existence d'un important biais par défaut dans l'estimation du nombre de personnes non recensées établie à partir de la CVD de 1971. L'estimation de 1971, si elle n'était pas biaisée serait d'environ 3,8% et non de 1,9%. Mais pour que le pourcentage réel soit bel et bien de 3,8%, il faudrait qu'entre 1966 et 1971, l'accroissement du surdéveloppement n'ait pas été aussi important et qu'il y ait eu augmentation du surdéveloppement en 1976 puis diminution en 1981. Il faudrait également qu'il y ait eu sous-estimation des personnes non recensées en 1976.

Toutefois, un tel scénario est purement théorique et l'on n'a d'ailleurs pas réussi à déterminer pourquoi de tels changements se seraient produits. Il y a d'autres scénarios possibles. Il n'en demeure pas moins que l'existence de ces différences soulève des doutes sur la fiabilité des estimations de la CVD et sur les effets possibles du surdéveloppement sur l'erreur nette de couverture.

La distribution provinciale de la différence entre l'estimation du nombre de personnes recensées établie au moyen de la CVD et le chiffre du recensement varie selon le recensement et rend encore plus difficile l'évaluation des effets de cette différence et ses causes possibles. Les résultats, pour 1976, sont présentés au tableau 5.

3.2 Variation dans les distributions géographiques

Les estimations du sous-développement de la CVD peuvent servir d'indicateurs généraux de la qualité des chiffres du recensement. Ces estimations doivent par ailleurs être utilisées dans l'élaboration et l'essai de méthodes visant à améliorer la couverture du recensement. Idéalement, on voudrait pouvoir utiliser les estimations de la CVD pour modéliser le sous-développement et produire par ce moyen des estimations régionales qui serviraient pour le redressement des chiffres du recensement. Toutefois, les différences géographiques dans la

Différence entre le nombre de personnes recensées établi à partir de la CVD et le chiffre de population du recensement de 1976

Province	Différence entre le nombre de personnes recensées (CVD - chiffre du recensement)	Différence en pourcentage
Canada (10 provinces)	-323,500	-1.4
Terre-Neuve	21,900	3.9
Ile-du-Prince-Édouard	-500	-0.4
Nouvelle-Écosse	-4,500	-0.5
Nouveau-Brunswick	-15,000	-2.3
Québec	-56,200	-0.9
Ontario	-207,000	-2.5
Manitoba	-6,600	-0.6
Saskatchewan	1,400	0.1
Alberta	-43,400	-2.4
Colombie-Britannique	-12,800	-0.5

Comme le fait voir le tableau 4, le nombre de migrants pendant la période intercensitaire a été sous-estimé pour toutes les provinces sauf la Saskatchewan. Cela peut être causé en partie par les cas «non dépités». Pour la Colombie-Britannique, cette sous-estimation peut expliquer la différence figurant au tableau 3 pour cette province. Par contre, la sous-estimation des migrants en Alberta n'explique pas adéquatement la différence calculée pour cette province, et cette différence doit donc tenir à l'un ou à plusieurs des facteurs énumérés aux points (i) à (v) ci-dessus.

La sous-estimation des migrants peut entraîner une distorsion dans les estimations du sous-dénombrement à l'intérieur des provinces, c'est-à-dire que les différences considérables entre les provinces (voir tableau 4) peuvent être causées par un biais très important dans les taux de sous-dénombrement provinciaux. De plus, comme nous l'avons souligné dans la section 2.2.3, le taux de sous-dénombrement des migrants est supérieur au taux moyen. Si le nombre de personnes recensées chez les migrants est sous-estimé, tandis qu'en général le nombre de non-migrants ne l'est pas (par rapport au chiffre du recensement), alors les estimations du sous-dénombrement peuvent être trop basses.

Des différences ont également été observées, dans les recensements antérieurs, entre le nombre estimatif de personnes recensées établi à partir de la CVD et le chiffre de population du recensement. Au recensement de 1971, la différence entre l'estimation de la CVD et le chiffre du recensement était de 289,000 et, au recensement de 1976, de -324,000. Pour ces deux recensements, les estimations de la CVD relatives aux personnes décédées, émigrées et à l'étranger étaient comparables aux estimations provenant d'autres sources. Le changement considérable observé en 1976 par rapport à 1971, conjugué aux valeurs négatives élevées observées

Tableau 4

CVD: Estimation du nombre de migrants¹
recensés en 1981, par province

Province	Estimation		Recensement: estimation de la migration interprovinciale	
	CVD	Chiffre officiel du recensement ²	Différence: CVD-recensement	Entrées
Canada	4,670,311	5,046,500	-376,239	1,124,970
Terre-Neuve	61,499	72,100	-10,601	18,430
Ile-du-Prince-Édouard	13,257	20,530	-7,273	9,945
Nouvelle-Écosse	125,949	137,865	-11,916	54,455
Nouveau-Brunswick	96,607	109,955	-13,348	41,460
Québec	1,092,919	1,145,085	-52,166	61,310
Ontario	1,572,504	1,725,225	-152,721	250,570
Manitoba	143,391	165,105	-21,714	54,030
Saskatchewan	204,937	192,840	12,097	63,395
Alberta	669,995	691,970	-21,975	336,830
Colombie-Britannique	689,253	785,825	-96,622	234,545
				123,615
				0.53

¹ Un migrant est une personne qui au moment du recensement précédent vivait à l'extérieur du Canada, dans une autre province ou une autre municipalité (ou une autre SDR). Les données sur la mobilité présentées ci-dessus sont celles qui ont été fournies au recensement par les personnes sélectionnées dans l'échantillon de la CVD et non celles obtenues au moment de la vérification des adresses.

² Statistique Canada 1983a.

personnes (67,000 plus 18,000) aient toutes démenagé en Alberta et en Colombie-Britannique pour que les différences pour ces provinces respectent un intervalle de confiance de 95%. C'est

la une supposition tout à fait déraisonnable.

Ce qui reste de la différence entre les estimations (125,000 personnes) peut être attribuable à diverses sources (potentielles) d'erreur dans la CVD ou dans le recensement: (i) erreur d'échantillonnage dans l'estimation des personnes recensées établie à partir de la CVD; (ii) accroissement du surdénombrement au recensement de 1981 par rapport à celui de 1976; (iii) exclusion, pour la CVD, de tous les immigrants illégaux et personnes revenant au statut de réfugié qui ont été dénombrés au recensement; (iv) sous-estimation des personnes «non recensées» en 1976, ces personnes formant dans la «base des personnes non recensées» pour la CVD de 1981; (v) surestimation du nombre de personnes «non recensées» en 1981. Cependant, on ne sait pas quelle part peut avoir chacune de ces sources d'erreur dans la différence. Le fait qu'une partie importante de la différence semble être attribuable à la Colombie-Britannique et à l'Alberta peut, dans une certaine mesure, s'expliquer par une sous-estimation de la migration inter-sitaire. La migration vers ces provinces a été particulièrement élevée entre 1976 et 1981 (Statistique Canada 1979; 1983a).

Il se peut également que les estimations des personnes décédées, émigrées ou à l'étranger soient biaisées. Si ces estimations sont en fait trop élevées, mais pour une raison autre qu'un biais associé aux «personnes non dépistées», il devrait également y avoir sous-estimation des personnes non recensées parce qu'aux fins de la CVD c'est la dernière adresse au Canada qu'on cherche à établir et c'est à partir de cette adresse que la recherche est effectuée. Les personnes qui ont émigré, qui sont mortes ou qui sont parties à l'étranger après le jour du recensement ont peut-être été inscrites comme telles au moment du dépistage, plusieurs mois après le jour du recensement. Mais en même temps, le fait qu'il ne semble pas y avoir sous-estimation du nombre de personnes décédées, et cela malgré les exclusions des bases de la CVD, semble indiquer que les personnes exclues (comme les immigrants, voir tableau 2) ont un taux de mortalité inférieur à celui du reste de la population et/ou que les estimations sont trop élevée pour ce groupe.

Tableau 3

CVD: Estimation du nombre de personnes recensées en 1981, par province

Province	CVD, nombre estimatif de personnes recensées	E.T. de l'estimation de la CVD	Chiffre officiel du recensement ¹	Nombre de personnes recensées, CVD - Recensement	CVD, nombre estimatif de personnes recensées
Canada (10 provinces)	24,064,376	62,193	24,274,287	-209,912	497,277
Terre-Neuve	568,696	8,256	567,681	1,015	10,039
Ile-du-Prince-Edouard	116,012	3,005	122,506	-6,494	1,456
Nouvelle-Ecosse	837,045	11,185	847,442	-10,397	9,034
Nouveau-Brunswick	685,332	8,167	696,403	-11,071	12,864
Québec	6,410,662	38,648	6,438,403	-27,736	125,180
Ontario	8,629,374	52,802	8,625,107	4,267	171,010
Manitoba	1,028,162	15,133	1,026,241	1,921	10,203
Saskatchewan	973,450	11,740	968,313	5,137	9,712
Alberta	2,151,480	24,238	2,237,724	-86,242	58,335
Colombie-Britannique	2,664,163	19,798	2,744,467	-80,304	89,445

¹ Statistique Canada 1982.
² Plus de trois erreurs types.

3. RÉSULTATS DE LA CVD

La CVD permet d'établir non seulement des estimations du nombre de personnes n'ayant pas été dénombrées au recensement, mais aussi des estimations indépendantes du nombre de personnes recensées ainsi que des personnes qui sont décédées, ont émigré ou ont séjourné à l'étranger pendant la période intercensitaire. Ces estimations indépendantes servent à valider les estimations du sous-dénombrement de CVD. Certains résultats de cette validation permettent d'illustrer les lacunes dont nous avons parlé dans la section 2.

Des analyses ont en outre été effectuées pour établir la relation entre les variations géographiques dans le sous-dénombrement et les variations dans la distribution des caractéristiques de la population et des ménages.

3.1 Estimations indépendantes

Les estimations de la CVD relatives au nombre de personnes recensées et de personnes décédées ainsi qu'au nombre de personnes ayant quitté le Canada durant la période intercensitaire peuvent être comparées aux estimations provenant d'autres sources. Par exemple, les estimations de la CVD sur le nombre de personnes recensées peuvent être comparées aux chiffres du recensement et les estimations des personnes décédées, aux statistiques de l'état civil. Si les estimations établies au moyen de la CVD ne comportent pas de biais significatifs, l'écart entre les estimations poura généralement être expliqué par l'erreur d'échantillonnage correspondant à l'estimation de la CVD. Si l'écart entre les estimations est important, il pourrait alors être attribuable à un biais dans les estimations de la CVD. La qualité globale des estimations de la CVD, qu'on évalue en faisant ces comparaisons, reflètera vraisemblablement la qualité des estimations des «personnes non recensées».

Les estimations de la CVD en ce qui concerne les émigrants (296,727) et les personnes «à l'étranger» (57,909) se comparent favorablement aux estimations établies au moyen de l'analyse démographique. Par exemple, sur les cinq valeurs établies par analyse démographique, les estimations de la CVD pour les émigrants se situaient à mi-chemin entre les valeurs extrêmes (197,000 et 372,000, la médiane étant de 266,400). L'estimation de la CVD en ce qui concerne les personnes décédées (846,378) est très proche du chiffre publié par Statistique Canada (840,589) pour la période 1976-1981.

Cependant, la comparaison des estimations des personnes recensées met en évidence des écarts importants. Certaines de ces estimations sont présentées au tableau 3 et d'autres, à l'annexe 1. Pour l'ensemble du Canada (10 provinces) et pour deux des dix provinces, le nombre de personnes recensées estimé au moyen de la CVD est très différent du chiffre officiel du recensement. L'écart de 209,911 personnes calculé pour l'ensemble du Canada peut en partie être expliqué par les exclusions des bases de la CVD. Mais les écarts entre les provinces sont difficiles à expliquer, et c'est ce qui les rend importants. Relativement à l'écart de 209,907 au niveau national, il faut tenir compte du fait que l'estimation du nombre de personnes non recensées établie au moyen de la CVD est de 497,277. De même, l'écart de 80,304 pour la Colombie-Britannique et de 86,244 pour l'Alberta doit être examiné en tenant compte du fait que le nombre estimatif de personnes non recensées établi par la CVD pour l'une et l'autre province est de 89,445 et de 58,335 respectivement.

On estime à 67,000 le nombre de ressortissants canadiens qui étaient à l'étranger au moment du recensement de 1976 et qui sont rentrés au Canada avant le recensement de 1981, et à 18,000 le nombre de personnes qui ont quitté un territoire pour s'établir dans une province pendant la période intercensitaire. Si l'on suppose qu'aucune de ces personnes n'a été oubliée au recensement, la différence entre les estimations du recensement et celles de la CVD ne serait plus que d'environ 125,000 personnes. Cette différence est encore trop considérable pour qu'on puisse l'attribuer uniquement à l'erreur d'échantillonnage. De plus, il aurait fallu que ces 85,000

On pourrait constituer un échantillon de personnes «à l'étranger» à partir de la CVD précédente. Cet échantillon serait toutefois de très petite taille et ne serait pas représentatif de l'ensemble du groupe en question. Le dépistage des personnes sélectionnées serait en outre difficile à effectuer.

Hors les immigrants illégaux, le groupe des personnes qui n'ont jamais été recensées deviendra de moins en moins nombreux. Les immigrants illégaux arrivés pendant la période intercensitaire et les autres immigrants illégaux n'ayant jamais été recensés au Canada demeureront exclus.

On pourrait réduire les effets de l'erreur d'échantillonnage en augmentant la taille de l'échantillon. Mais la taille serait augmentée de combien, à quel coût et en fonction de quels critères? Si l'on faisait passer l'échantillon de la CVD de 36,550 à 100,000 personnes, cela serait probablement suffisant pour ramener à moins de 0,2% les estimations provinciales de l'erreur type pour les taux de sous-dénombrement. Toutefois, suivant les estimations du sous-dénombrement qu'on obtiendrait en fait, un échantillon de cette taille ne serait peut-être pas assez gros pour permettre le redressement des chiffres du recensement. Pour réduire l'erreur type à 0,1% pour chaque province – niveau produit par la CVD de 1981 et de 1976 pour un taux de sous-dénombrement estimatif de 2% pour l'ensemble du Canada (10 provinces) – il faudrait un échantillon de 350,000 personnes, en supposant les mêmes taux provinciaux de sous-dénombrement, le même genre de plan de sondage et les mêmes effets du plan. Effectuer une CVD de bonne qualité pour un si gros échantillon, étant donné les opérations de contrôle et de vérification de la qualité nécessaires, serait beaucoup plus coûteux que ne porterait à le penser la seule augmentation de la taille, et une telle mesure pourrait être irréalisable sur le plan opérationnel. Bien sûr, l'augmentation de la taille de l'échantillon ne permettrait pas de réduire le biais dans les estimations.

Les méthodes de dépistage employées dans la CVD font l'objet d'un examen avant et après chaque CVD. Des modifications majeures ont été faites en 1986 et on prévoit d'autres améliorations pour 1991. On doit toutefois s'attendre qu'il y ait encore un pourcentage non négligeable de personnes «non dépistées», pour lesquelles on continuera de recourir à la pondération, ou à l'imputation et la pondération.

Des évaluations peuvent être effectuées afin de déterminer la qualité de l'appariement et l'exactitude des adresses fournies par les répondants ou d'autres sources fiables. On pourrait également évaluer, dans une certaine mesure, l'effet potentiel de l'algorithme et des critères d'appariement. Toutefois, même si ces études permettaient de déceler un problème, elles n'offriraient pas nécessairement de solution.

On peut mettre à l'essai des modifications aux méthodes de pondération afin de mieux tenir compte de la mobilité et d'autres caractéristiques lorsqu'on effectue la correction de poids pour les «personnes non dépistées» (Burgess 1986). Les données supplémentaires requises pourraient peut-être être tirées des dossiers administratifs. Certains redressements mineurs pourraient également être faits à partir des renseignements dont on dispose en ce moment. Par exemple, l'ajustement pour les «personnes non dépistées» avec lesquelles un contact a pu être établi mais pour lesquelles on n'a pas réussi à obtenir l'adresse le jour du recensement pourrait être différent de l'ajustement effectué pour tenir compte des personnes non dépistées qui pouvaient être décédées, émigrées ou à l'étranger.

On pourrait également mettre à l'essai les corrections effectuées à l'aide des chiffres de population du dernier recensement. Toutefois, pour que ces corrections puissent réduire le biais associé aux «personnes non dépistées» et aux personnes qui ne sont pas incluses dans l'échantillon de la CVD, il faudrait que la classification de base des «personnes non recensées» ne comporte pas de biais et qu'il n'y ait pas non plus de distorsion interprovinciale de la proportion des «personnes non recensées». Ces changements aux méthodes de pondération ne permettraient pas, en elles-mêmes, d'éliminer le biais.

(personnes dépitées) dans les différents groupes de pondération. La façon dont est effectué l'ajustement pour tenir compte des personnes non dépitées dépend de la quantité de renseignements dont on dispose sur ces personnes. Idéalement, il faudrait, lorsqu'on définit les groupes de pondération, tenir compte des facteurs suivants: comment les personnes choisies ont-elles été dépitées? ont-elles démenagé et, le cas échéant, à quel endroit? quelles sont leurs caractéristiques démographiques? Jusqu'à maintenant, on ne s'est servi que des caractéristiques démographiques et d'un minimum de données sur la mobilité pour effectuer la correction des poids. (Cette correction n'est pas appliquée aux personnes sélectionnées dans la «base du recensement» qui n'ont pas démenagé pendant la période intercensitaire et qui sont classées comme «recensées»). Dans l'état actuel des choses, il serait difficile de classer la majorité des «personnes non dépitées» car à leur sujet on sait seulement qu'elles n'ont pas été recensées à l'adresse indiquée sur la liste.

Pour la seconde correction de poids, on obtient pour chaque base (sauf pour la base des «personnes non recensées», qui est un échantillon) les chiffres de population à l'intérieur des différents sous-groupes considérés. À l'aide de ces totaux «connus», on effectue une correction des poids de la CVD à l'intérieur des sous-groupes correspondants de l'échantillon. Cela permet de réduire l'erreur dans les estimations établies par la CVD en s'assurant que les totaux obtenus pour l'échantillon, eu égard à certaines caractéristiques démographiques de base pour lesquelles des taux de sous-dénombrement sont publiés, correspondent aux totaux obtenus pour les bases.

Ni l'une ni l'autre de ces corrections de poids ne permet de tenir compte des différentes exclusions.

L'estimateur prend la forme générale suivante:

Proportion estimée de personnes non recensées

$$= \frac{\text{Nombre estimé de personnes non recensées}}{\text{Nombre de personnes recensées} + \text{Nombre estimé de personnes non recensées}}$$

Cet estimateur est examiné de façon plus approfondie à l'annexe 2.

2.3 Réduction des sources d'erreur et des lacunes de la CVD

Des études expérimentales et l'évaluation de la méthodologie de la CVD permettraient peut-être d'éliminer certaines lacunes ou sources d'erreur, ou d'en réduire les effets. Une étude indépendante permettrait peut-être d'estimer le surdénombrement. On effectue actuellement, à titre expérimental, une telle étude pour le recensement de 1986. Toutefois, la production d'estimations provinciales de bonne qualité pourrait être très coûteuse.

L'établissement d'estimations pour le Yukon et les Territoires du Nord-Ouest nécessiterait la création d'un ensemble de listes ou de bases autres que celles utilisées pour la CVD. Ces listes devraient être à jour et ne devraient pas comporter de doubles comptes qu'on ne serait pas en mesure de supprimer ou d'estimer. À l'aide de cet ensemble de listes, il serait possible d'appliquer aux territoires la méthodologie de base de la CVD. Certains travaux expérimentaux ont déjà été réalisés à ce chapitre et on en prévoit d'autres.

On pourrait grossir les listes servant à la CVD afin d'y inclure certaines catégories de personnes exclues, par exemple les personnes revendiquant le statut de réfugié et celles ayant quitté les territoires pour s'établir dans une province. Toutefois, le dépistage de ces personnes serait difficile. Par ailleurs, il se peut que l'échantillonnage de ces groupes ait uniquement pour résultat de changer la nature du problème.

Certains cas de surdénumérement dans les bases peuvent être détectés, par exemple résidents étrangers dénumérés au recensement précédent, noms « créés » à la suite d'une erreur de traitement au recensement précédent, immigrants qui n'étaient pas encore établis au Canada, enfants nés au Canada de parents non résidents, personnes qui n'existent pas ou ne font pas partie de l'univers du recensement mais ont été inscrites dans un questionnaire au recensement précédent. En 1981, ces cas représentaient moins de 0,1 % des personnes sélectionnées. D'autres cas de surdénumérement dans les bases ne peuvent pas être détectés, en particulier les doubles comptes dans une base. Il se peut en outre qu'on ne se rende pas compte qu'une personne sélectionnée n'existe pas et qu'on la classe dans la catégorie des personnes « non dépistées ».

Le classement final des personnes sélectionnées pour la CVD de 1981 est présenté au tableau 2 (tiré de Burgess 1986).

2.2.5 Pondération et estimation

Au moment de la sélection de l'échantillon, on a attribué à chaque personne sélectionnée un poids correspondant à l'inverse de la fraction de sondage. Deux types de corrections ont été apportées à ce poids: l'une pour tenir compte des personnes «non dépistés», l'autre pour compenser le fait que la suppression des cas non dépistés donne une moins bonne représentativité à l'échantillon par rapport aux bases dont il est tiré.

Les personnes «non dépistées» correspondent aux cas suivants: personnes qui ont été recensées ou non au moment du recensement, personnes décédées avant le recensement, qui ont émigré ou qui étaient à l'étranger, cas de surénumérement. Les poids calculés pour tenir compte des personnes «non dépistées» sont donc redistribués parmi les «personnes dépistées». Cette correction se fait à l'intérieur de groupes définis en fonction de la base et de différentes caractéristiques démographiques et géographiques.

La correction de poids pour tenir compte des «personnes non dépistées» se fait en deux temps. Un premier ajustement est fait pour tenir compte des personnes sélectionnées pour lesquelles aucun dépistage n'a été effectué parce qu'on ne disposait pas des renseignements nécessaires pour faire une recherche et en arriver à un appariement et il est appliqué à toutes les autres personnes sélectionnées (personnes non dépistées qui restent et personnes dépistées). Un second ajustement est fait pour tenir compte des personnes sélectionnées pour lesquelles un dépistage a été fait mais sans succès. Cet ajustement est appliqué aux personnes qui restent

Tableau 2
CVD de 1981: classement final des personnes sélectionnées

Classement final	Base			
	Recensement	Naissances	Immigrants	Non recens.
	Nombre %	Nombre %	Nombre %	Nombre %
	Total			

Personnes dépitées										
29,761	97.1	3,211	92.3	1,392	96.1	807	96.1	35,171	96.6	96.6
Recensées	27,541	89.8	3,096	89.0	1,113	76.8	696	82.9	32,446	89.1
Décédées	1,056	3.5	33	0.9	5	0.3	26	3.1	1,120	3.1
Emigrées/à l'étranger	299	1.0	34	1.0	111	7.7	24	2.8	468	1.3
Non recensées	865	2.8	48	1.4	163	11.3	61	7.3	1,137	3.1
Personnes non										
dépitées (y compris										
les cas de surdénom-										
brement)										
895	2.9	267	7.7	57	3.9	33	3.9	1,252	3.4	3.4
30,656	100.0	3,478	100.0	1,449	100.0	840	100.0	36,423	100.0	100.0
TOTAL										

données d'identification dans le questionnaire du recensement et dans les documents utilisés pour le CVD ne contiennent pas d'erreurs (c'est-à-dire qu'il n'y ait pas eu d'erreur de réponse ou non-réponse en ce qui concerne ces éléments d'information).

Si les données d'identification de la personne choisie sont justes (nom complet, âge, sexe, etc.) et qu'il n'y pas eu d'erreur de traitement, aucune personne choisie qui n'a pas été recensée ne sera classée dans la catégorie des «personnes recensées». Toutefois, l'inverse n'est pas vrai. Si la personne sélectionnée a été recensée à une adresse autre que celle fournie dans la base (ou tirée des dossiers administratifs ou de l'annuaire téléphonique), il faut communiquer avec elle (ou avec une source fiable) pour obtenir cette adresse. Si la personne rejointe ne donne pas la bonne adresse (par exemple, se trompe ou ne se souvient pas), la personne sélectionnée sera classée dans la catégorie des «personnes non dépistées» ou des «personnes non recensées». Habituellement, quand la personne sélectionnée (ou un membre de la famille, le conjoint ou une autre source fiable) fournit l'adresse (ou les adresses) à laquelle elle aurait dû ou peut avoir été recensée, cette adresse est considérée comme correcte. Les personnes sélectionnées sont classées dans la catégorie des «personnes recensées» ou des «personnes non recensées» sur la foi des renseignements qu'on aura obtenus sur l'adresse le jour du recensement. Lorsqu'on classe quelqu'un dans la catégorie des «personnes non recensées», on ne peut pas savoir si l'adresse est en fait bonne ou mauvaise.

En outre, la probabilité qu'une personne non recensée soit classée dans la catégorie des «personnes non dépistées» peut être plus grande que dans le cas d'une personne qui a bel et bien été recensée. Avant qu'une personne sélectionnée puisse être classée dans la catégorie des «personnes non recensées», il faut communiquer avec elle (ou avec une source fiable) pour lui demander de confirmer l'adresse qu'on a entre les mains ou d'indiquer l'adresse (ou les adresses) à laquelle elle a pu être recensée et de fournir, pour elle-même et les membres de son ménage, certaines données du recensement. Cette façon de procéder permet d'éliminer un certain nombre d'erreurs de classement. Toutefois, si l'on a des doutes quant à l'exactitude des renseignements relatifs à une personne non recensée, il faut communiquer avec elle (ou avec un membre de sa famille, son conjoint, etc.) pour obtenir des précisions. Si les doutes ne sont pas dissipés, la personne en question sera classée dans la catégorie des personnes «non dépistées». Dans le cas des personnes recensées, il n'est pas toujours nécessaire que les renseignements soient aussi concluants. En poursuivant le dépistage, il se peut qu'on arrive à reclasser comme ayant été recensée une personne qui a été comptée dans la catégorie des «personnes non dépistées», même si l'adresse obtenue est incomplète ou incorrecte. Un tel dépistage ne changerait rien dans le cas d'une personne non recensée.

Les données d'identification de la personne sélectionnée ne sont pas toujours justes et complètes. Il arrive parfois, lorsqu'on effectue l'appariement, qu'on classe la personne sélectionnée dans la catégorie des «personnes recensées» même si le nom indiqué dans le questionnaire et celui figurant dans les dossiers de la CVD ne sont pas rigoureusement identiques. Parfois, un nom complet n'est fourni que pour la première personne inscrite dans le questionnaire et, dans d'autres cas, il n'y a pas de nom du tout. (Si l'on n'arrive pas à établir l'identité de la personne sélectionnée à partir de la base, on va tout de suite la classer dans la catégorie des personnes «non-dépistées». Cela sera fait notamment dans les cas où le ménage était absent au recensement précédant et où le nombre de personnes comprises dans celui-ci aura été attribué par Statistique Canada ou encore dans les cas où le ménage aura refusé de participer au recensement précédant). Pour la date de naissance et les autres données, il peut y avoir des lacunes ou encore des différences entre la base et le questionnaire. L'exactitude de l'appariement est assurée dans la majorité des cas mais pas toujours. Certains cas considérés comme appariés seront classés à tort dans la catégorie des «personnes recensées». Des cas rejetés à l'appariement seront classés par erreur comme «non recensés», encore que la plupart seront classés comme «non retracés». D'autres règles d'acceptation et de rejet à l'appariement pourraient, bien entendu, produire des estimations de sous-dénombrement différentes.

fois la proportion initiale (pondérée) de personnes non recensées par rapport à l'ensemble des personnes sélectionnées qui ont été dépitistes. Il pourrait donc exister un lien entre les personnes «non dépitistes» et les «personnes non recensées». On ne sait pas, bien sûr, si le taux de 1.6 est trop élevé, trop bas ou exact. S'il n'est pas exact, il pourrait y avoir distorsion dans les estimations provinciales du sous-dénombrement et un biais dans les estimations globales du sous-dénombrement.

Comme le taux de migration interprovinciale entre deux recensements varie d'une province à l'autre, il se peut qu'il y ait une certaine distorsion dans les estimations provinciales si la proportion des migrants interprovinciaux dans les groupes de pondération n'est pas la même chez les personnes «dépitistes» et «non dépitistes».

Le taux de sous-dénombrement est élevé chez les migrants interprovinciaux («base du recensement» et «base des personnes non recensées» seulement). On a estimé ce taux à 6.13% pour le recensement de 1981; on a pour cela utilisé les données sur la mobilité de la CVD de 1981, elles-mêmes obtenues par comparaison des adresses aux recensements de 1976 et 1981. Le taux de sous-dénombrement estimé pour les personnes ayant migré à l'intérieur d'une province pendant la période intercensitaire (personnes ayant déménagé dans une autre subdivision de recensement, SDR, ou une autre municipalité) était de 3.83%. Chez les personnes ayant déménagé à l'intérieur d'une SDR ou d'une municipalité, le taux de sous-dénombrement était estimé à 2.83%. Compte tenu de ces taux et de la distribution des caractéristiques de mobilité, le taux de sous-dénombrement «imputé» pour tenir compte des personnes «non dépitistes» dans la «base des personnes recensées» et la «base des personnes non recensées» (prises ensemble) devrait être d'au moins 3.52%, non de 3.27%, étant donné que les personnes «non dépitistes» sont presque toujours des personnes qui ont déménagé. Par ailleurs, on peut s'attendre que les «personnes non dépitistes» comprennent proportionnellement autant de personnes ayant migré à l'intérieur d'une province et d'une autre et qu'elles aient des taux de sous-dénombrement, selon la mobilité, supérieurs ou égaux à ceux des personnes dépitistes. (D'après la CVD de 1981, la distribution estimative (en pourcentage) de la population âgée de 5 ans et plus, en ce qui concerne la mobilité, était la suivante: (i) personnes n'ayant pas déménagé, 55%; (ii) personnes ayant déménagé à l'intérieur d'une SDR, 17%; (iii) personnes ayant déménagé dans une autre SDR dans la même province, 21.7%; (iv) personnes ayant déménagé dans une autre province, 5%; et (v) personnes ayant immigré au Canada, 2%.

Etant donné les méthodes de dépitage utilisées, il n'est pas déraisonnable de supposer que la proportion de migrants et, par le fait même, le taux de sous-dénombrement sont beaucoup plus élevés chez les personnes «non dépitistes». Si tel est le cas, les estimations du sous-dénombrement pourraient être biaisées vers le bas de façon significative. Si, par exemple, le «vrai» taux de sous-dénombrement des «personnes non dépitistes» était d'environ 5%, le biais dans l'estimation du sous-dénombrement pour l'ensemble du Canada (10 provinces) serait plus élevé que l'erreur d'échantillonnage.

2.2.4 Recherche et classification

Après avoir terminé le dépitage et effectué les interviews nécessaires, chaque personne sélectionnée est classée dans une des six catégories suivantes:

- (1) personnes recensées,
- (2) personnes non recensées,
- (3) personnes dépitistes,
- (4) personnes ayant émigré ou personnes qui séjourneraient à l'étranger,
- (5) cas de surdénombrement dans une base,
- (6) personnes non dépitistes.

Comme nous l'avons souligné, afin de déterminer si une personne sélectionnée a été recensée ou non, il faut trouver le questionnaire du recensement correspondant à son adresse. Pour qu'on puisse classer la personne dans la bonne catégorie, il faut que l'adresse soit bonne et que les

dénombrement est estimé à 2%, le coefficient de variation serait d'à peu près 50%. Dans le cas de petites régions et de petites populations, le coefficient de variation pourrait être beaucoup plus grand encore.

Bien entendu, l'erreur d'échantillonnage n'est pas sans effet sur la possibilité d'observer les différences de taux de sous-dénombrement entre des provinces et entre d'autres unités géographiques. Il devient alors difficile de déterminer les causes du sous-dénombrement ainsi que les secteurs géographiques et segments de la population à l'intérieur desquels il se manifeste, et de compenser l'erreur de couverture pour améliorer les chiffres du recensement. Les provinces dont le taux de sous-dénombrement est nettement différent de celui d'une ou de plusieurs autres figures également au tableau 1. Pour le recensement de 1981, les taux de sous-dénombrement provinciaux et l'écart entre ces taux permettent de classer les provinces en six groupes. Au recensement de 1976, où l'on a défini huit groupes, il y avait davantage de différence entre les provinces. Toutefois, dans ni l'un ni l'autre recensement on ne peut montrer qu'aucun de ces groupes était tout à fait différent de tous les autres; et peut-être était-ce en fait le cas.

Ces résultats ne sont pas très différents de ceux de la CVD pour les recensements de 1966 et de 1971. Depuis 1966, seule la Colombie-Britannique a eu un taux de sous-dénombrement sensiblement plus élevé que celui de l'ensemble du Canada. Pour la plupart des provinces, la variation du taux de sous-dénombrement d'un recensement à l'autre peut être attribuable, dans une large mesure, à l'erreur d'échantillonnage. Mais ce n'est pas le cas pour la Colombie-Britannique, et c'est là un sujet de vive préoccupation en ce qui concerne la CVD et le recensement.

La nécessité d'utiliser un échantillon des personnes «non recensées» établi à partir de la CVD précédente est aussi une contrainte en ce qui concerne le plan de sondage et la taille de l'échantillon. Il n'y a aucune façon de fixer la taille de cette partie de l'échantillon, et les lacunes de la CVD précédente seront perpétuées dans la CVD suivante dans la mesure où l'estimation des personnes «non recensées» en était le reflet. (Voir sections 2.2.4, 2.2.5 et 3).

2.2.3 Dépistage

Étant donné la façon dont sont constituées les bases (ou listes) dont est tiré l'échantillon de la CVD, les adresses et autres données peuvent avoir jusqu'à cinq ans et n'être plus à jour. On essaie donc, à l'aide des dossiers administratifs, de mettre à jour les adresses avant le jour du recensement. (Des travaux exhaustifs de mise à jour ont été faits pour la première fois dans le cadre de la CVD de 1986). Après le jour du recensement, on commence par chercher les questionnaires du recensement correspondant à l'adresse originale ou à l'adresse corrigée (si on a pu l'obtenir) pour déterminer si la personne sélectionnée a été recensée. Si, après cette première recherche, on n'arrive pas à déterminer si la personne sélectionnée a été recensée, il faut continuer le dépistage, c'est-à-dire soit communiquer avec la personne choisie (ou avec une source fiable) pour obtenir l'adresse corrigée, soit déterminer le statut de la personne sélectionnée (décédée, émigrée, à l'étranger, etc.).

Bien que les opérations de dépistage soient exhaustives, on n'arrive pas toujours à retrouver toutes les personnes sélectionnées, ce qui peut engendrer une certaine forme de biais de non-réponse. À la CVD de 1981, 3,4% des personnes sélectionnées n'ont pu être dépistées. Le taux global de sous-dénombrement du recensement étant estimé à 2%, ce taux de personnes «non dépistées» est un des facteurs qui compromettent le plus la qualité des estimations produites à partir de la CVD.

Une correction de poids a été effectuée pour tenir compte des personnes «non dépistées». Au recensement de 1981, l'effet de cette correction a été l'attribution d'un taux de sous-dénombrement de 3,27% dans la «base du recensement» et la «base des personnes non recensées» (prises ensemble), de 1,46% pour la «base des naissances» et de 1,94% pour la «base des immigrants». Globalement, la proportion des poids calculés pour tenir compte des personnes non dépistées qui a été attribuée aux personnes non recensées correspondait à 1,6

recensement précédent qui n'auraient pas dû l'être ou qui ont été dénombrés deux fois, ou encore des personnes qui n'existent pas ou qui ont été comptées plus d'une fois au moment du traitement. La CVD permet de déceler une partie du surdénombrement. Toutefois, lorsqu'on estime le sous-dénombrement, l'effet du surdénombrement dans les bases ne devient important que si son taux est égal ou supérieur au taux de sous-dénombrement.

2.2.2 Plan d'échantillonnage et taille de l'échantillon

L'erreur d'échantillonnage est une des principales faiblesses des résultats de la CVD. La grandeur de l'erreur est fonction du plan de sondage et de la taille de l'échantillon. La taille de l'échantillon est l'élément le plus important puisque, comme les bases, elle limite les possibilités en ce qui concerne l'élaboration du plan de sondage.

Les principales estimations du sous-dénombrement établies au moyen de la CVD de 1981 et de 1976, ainsi que l'erreur type correspondante, sont présentées au tableau 1. Le coefficient de variation (l'erreur type divisée par le taux estimatif du sous-dénombrement) varie beaucoup, de 4,5% pour l'ensemble du Canada (10 provinces) jusqu'à 13,6% pour les régions (Atlantique, Québec, Ontario, Prairies et Colombie-Britannique) et 46% pour les provinces. Les coefficients de variation au niveau infraprovincial sont généralement plus élevés. Par exemple, dans un district électoral de taille moyenne (86,323 personnes en 1981), où le taux de sous-

Estimation du taux de sous-dénombrement de la population
aux recensements de 1981 et de 1976, par province, et provinces
ayant un taux significativement différent intervalle de confiance 95%)

Province	Sous-dénombrement de la population	
	Taux (%)	E.-T. (%)

Recensement de 1981 Canada (10 Provinces)		
1. Terre-Neuve	1.74	0.45
2. Île-du-Prince-Édouard	1.17	0.54
3. Nouvelle-Écosse	1.05	0.34
4. Nouveau-Brunswick	1.81	0.30
5. Québec	1.91	0.21
6. Ontario	1.94	0.14
7. Manitoba	0.98	0.35
8. Saskatchewan	0.99	0.37
9. Alberta	2.54	0.36
10. Colombie-Britannique	3.16	0.33
Toutes sauf 9		
Recensement de 1976 Canada (10 Provinces)		
1. Terre-Neuve	1.10	0.39
2. Île-du-Prince-Édouard	0.38	0.25
3. Nouvelle-Écosse	0.86	0.34
4. Nouveau-Brunswick	2.16	0.37
5. Québec	2.95	0.25
6. Ontario	1.52	0.17
7. Manitoba	1.07	0.33
8. Saskatchewan	1.33	0.34
9. Alberta	1.49	0.26
10. Colombie-Britannique	3.13	0.31
Toutes sauf 5		

- (iv) Personnes non recensées – Échantillon des personnes non dénombrées au dernier recensement, c'est-à-dire le groupe de personnes qui, d'après les résultats de la CVD précédente, n'ont pas été recensées (il n'existe pas de liste complète de ce groupe).

Ces listes ou bases sont censées comprendre toutes les personnes qui doivent avoir été dénombrées dans l'une des dix provinces au dernier recensement, sans double compte de personnes sélectionnées sur une même liste ou sur deux listes différentes.

Certaines personnes ne figurent toutefois pas sur ces listes. Il s'agit entre autres: (a) des immigrants illégaux arrivés durant la période intercensitaire et n'ayant jamais été recensés; (b) de certaines catégories de réfugiés; (c) de certains Canadiens qui étaient à l'étranger au moment du recensement précédent et qui sont revenus au pays avant le dernier recensement; (d) des personnes qui ont quitté les territoires pour s'installer dans une des provinces au cours de la période intercensitaire; (e) des personnes qui n'ont été dénombrées dans aucun des recensements depuis 1961, quand la CVD a commencée à être appliquée, mais qui étaient des résidents canadiens avant 1961.

On suppose, mais sans moyen de le confirmer, que le nombre de personnes dans la catégorie (e) est maintenant assez petit pour ne plus être significatif sur le plan statistique. Pour la catégorie (d), on estime qu'au recensement de 1981 elle comptait environ 18,000 personnes. La plupart demeuraient dans les territoires au moment du recensement précédent (1976) mais il est probable qu'un certain nombre d'entre elles figurent dans la «base des naissances» et la «base des immigrants».

La catégorie (c) comprend les Canadiens qui travaillaient, étudiaient ou voyageaient à l'étranger et qui n'ont pas conservé de «résidence habituelle» au Canada durant leur absence de même que, le cas échéant, les enfants de ces Canadiens nés à l'étranger. Cette catégorie (c) ne comprend pas les membres des Forces armées canadiennes, les employés des Affaires extérieures et autres fonctionnaires (et leur famille) qui demeuraient à l'étranger: ces personnes auront été incluses dans la «base du recensement» et la «base des personnes non recensées». Au recensement de 1981, la taille estimative de ce groupe de personnes «à l'étranger» était d'environ 67,000.

Les réfugiés et les immigrants illégaux doivent être recensés, en supposant qu'ils n'ont pas de résidence habituelle à l'extérieur du Canada et qu'ils ne sont pas titulaires d'un visa de travail ou d'étudiant. Aux fins de la CVD de 1981 et 1986, les personnes qui ont fait une demande à l'étranger et qui ont été reçues au Canada comme réfugiés ont été incluses dans la «base des immigrants». Les personnes ayant fait une demande après leur arrivée au Canada n'ont été incluses dans la «base des immigrants» que si elles ont obtenu le statut de réfugié. En avril 1985, il y avait 12,500 cas de demandes de résidence présentées après l'arrivée au Canada (voir Plaut 1985). Certains immigrants illégaux peuvent figurer dans la «base du recensement» ou même dans la «base des personnes non recensées». Souignons cependant qu'en raison des amnisties accordées par l'administration fédérale dans les années 1970 et 1980, un certain nombre d'immigrants illégaux figurent dans la «base des immigrants».

Les exclusions appliquées dans le cadre de la CVD sont importantes en ce sens que les personnes exclues n'auront pas été dénombrées au dernier recensement (1986). Comme le taux de sous-dénombrement de la «base des immigrants» tend à être élevé (8,5% en 1981 comparativement à 2,0% pour l'ensemble des bases), on peut raisonnablement prévoir un taux lui aussi élevé de sous-dénombrement pour les personnes qui demandent le statut de réfugié. Il se peut que la majorité des immigrants illégaux n'aient pas été recensés. Ces éléments peuvent expliquer une proportion importante du sous-dénombrement (environ 500,000 personnes en 1981) estimé au moyen de la CVD. Il se peut que les réfugiés et les immigrants illégaux soient groupés dans quelques villes de certaines provinces, ce qui viendrait accroître l'effet des exclusions sur la fiabilité des estimations.

Par ailleurs, on peut s'attendre que les bases comprennent un certain nombre de personnes qui ne devraient pas y figurer (surdénombrement), c'est-à-dire des personnes dénombrées au

2.1 Applicabilité des estimations fournies par la CVD

Il y a des lacunes peuvent prendre plusieurs formes: différences entre les besoins des utilisateurs au niveau conceptuel et les éléments que la CVD tente de mesurer; défauts de conception qui empêchent la CVD de produire les résultats qu'on en attend; erreurs d'échantillonnage ou de réponse ou autres types d'erreurs. En modifiant des éléments précis de la CVD, il est possible d'éliminer ou de réduire certaines de ces lacunes. Mais il y en a d'autres qui vont persister ou qui, de par leur nature même, ne pourront pas être corrigées.

L'objet de la CVD est de produire, pour chacune des dix provinces, des estimations du sous-dénombrement dans le recensement. Cette méthode ne permet pas de produire des estimations de l'erreur de couverture nette. Notons par ailleurs qu'elle n'est appliquée aux chiffres des Territoires du Nord-Ouest et du Yukon.

La CVD permet d'estimer le pourcentage de la population omise au recensement, c'est-à-dire les personnes qui n'ont pas été recensées mais auraient dû l'être. Elle ne sert pas à estimer le surdénombrement (personnes recensées plus d'une fois et personnes recensées qui n'auraient pas dû l'être ou qui n'existent pas). On ne peut donc pas l'utiliser pour estimer l'erreur de couverture nette (sous-dénombrement moins surdénombrement). L'effet du surdénombrement, même si le chiffre est peu élevé, est fonction de sa distribution par rapport au sous-dénombrement. Par exemple, un taux de surdénombrement de 0.2% (un dixième du niveau de sous-dénombrement en 1976 et 1981) serait très important si le pourcentage, pour une province donnée, atteignait 0.5%.

La CVD n'a pas été appliquée aux deux territoires parce que la population y est petite et que les taux d'immigration et d'émigration internes (calculés entre deux années de recensement) y sont très élevés. Pour produire des estimations fiables ou, si l'on veut, pour réduire le plus possible l'erreur d'échantillonnage, il faudrait que l'échantillon des territoires soit proportionnellement assez grand (taux d'échantillonnage de 5%, ce qui donnerait un échantillon de 3,750 personnes). Dans les territoires, les taux d'immigration et d'émigration internes sont de 33.3% ou plus. Par conséquent, en supposant le tirage d'un échantillon proportionnel, il faudrait que 1,250 des 3,750 unités (en moyenne) de l'échantillon minimum requis soient des immigrants internes. Lorsqu'on effectue la CVD, on utilise des listes qui contiennent l'adresse de la majorité des résidents au dernier recensement, cinq ans plus tôt. L'adresse des immigrants internes ne peut être établie qu'au moment où est effectuée la CVD, ce qui, en soi, ne pose pas de difficulté. Toutefois, le taux d'échantillonnage fixé pour la CVD est de 0.15% seulement. Par conséquent, dans les territoires, la proportion des immigrants internes échantillonnés dans le cadre de la CVD serait vraisemblablement de 0.15%, pas de 5%, et l'échantillon ne comprendrait que 30 immigrants internes. Dans les limites actuelles de la CVD et sans engager de dépenses considérables, il est impossible de constituer un échantillon représentatif de la proportion de 33.3% ou plus. Par conséquent, le taux d'échantillonnage fixé pour la CVD est de 0.15% seulement. Par conséquent, dans les territoires, la proportion des immigrants internes échantillonnés dans le cadre de la CVD serait vraisemblablement de 0.15%, pas de 5%, et l'échantillon ne comprendrait que 30 immigrants internes. Dans les limites actuelles de la CVD et sans engager de dépenses considérables, il est impossible de constituer un échantillon représentatif de la proportion de 33.3% ou plus de la population territoriale constituée d'immigrants internes.

2.2 Méthodologie de la CVD

Chacune des cinq grandes opérations de la CVD est une source avérée ou potentielle d'erreur.

2.2.1 Base d'échantillonnage

L'échantillon utilisé pour la CVD est constitué à partir de quatre listes ou bases:

- (i) Recensement – Toutes les personnes dénombrées au recensement précédent; pour la CVD de 1986, les personnes recensées en 1981.
- (ii) Naissances – Toutes les naissances survenues durant la période intercensitaire (base cons-
- (iii) Immigrants – Tous les immigrants reçus pendant la période intercensitaire (base cons-
- (iv) Immigrants – Tous les immigrants reçus pendant la période intercensitaire (base cons-

tituée à partir des dossiers d'Emploi et Immigration Canada).

couverture pour les recensements futurs (évaluation des principales causes de l'erreur de couverture et des principaux secteurs géographiques et segments de la population dans lesquels elle se manifeste).

Depuis 1961, Statistique Canada produit et publie des estimations du sous-dénombrement pour chaque recensement de la population. À cette fin, le Bureau utilise la méthode dite de la contre-vérification des dossiers (CVD), qui comporte cinq grandes opérations ou étapes:

- (i) Établissement de bases – Dresser une série de listes distinctes dans lesquelles sont représentées toutes les unités de la population devant avoir été dénombrées au recensement.
- (ii) Plan de sondage et tirage de l'échantillon – Tirer un échantillon aléatoire parmi les personnes inscrites sur ces listes.
- (iii) Dépistage – Déterminer, pour chaque personne sélectionnée, l'adresse du domicile habituel le jour du recensement (ou vérifier si la personne est décédée ou a émigré avant le recensement).
- (iv) Recherche – Examiner les questionnaires de recensement remplis afin de déterminer si les personnes sélectionnées ont été recensées ou non.
- (v) Pondération et estimation – Pondérer les chiffres obtenus à partir de l'échantillon afin d'établir une estimation du nombre de personnes non recensées.

Une description plus détaillée de cette méthode est présentée dans Gosselin (1976) et dans Statistique Canada (1984).

D'autres méthodes, par exemple le dénombrement postcensitaire, l'analyse démographique et la vérification des dossiers administratifs, pourraient servir à estimer le sous-dénombrement. Toutefois, dans le contexte canadien, les résultats que l'on obtiendrait avec chacune de ces méthodes seraient probablement moins fiables que ceux établis à partir de la CVD. Les dénombrements postcensitaires omettent souvent les mêmes personnes ou ménages que ceux qui ont été omis au recensement. Les analyses démographiques, fondées sur des modèles, ne produisent pas d'estimations fiables sur l'émigration, ne mesurent que la variation dans la couverture nette entre deux recensements, ne permettent de déterminer ni de causes ni de cas précis de sous-dénombrement, et entraînent, au niveau infranational, des erreurs dans les estimations de la migration interne. Les possibilités qu'offrent la vérification des dossiers administratifs sont limitées par l'absence d'un système administratif national ayant une couverture plus exhaustive que le recensement ou dont les erreurs de couverture diffèreraient de façon indépendante de celles du recensement, ce qui permettrait l'utilisation d'un fichier administratif incomplet. Même si un tel système existait, il ne constituerait qu'une variante de la CVD, à moins d'avoir été parfaitement à jour (couverture et adresses) au moment du recensement.

C'est pour ces raisons qu'au Canada on a retenu la CVD pour l'évaluation de la couverture, bien qu'on utilise aussi l'analyse démographique pour confirmer les résultats. Toutefois, la CVD comporte elle-même des lacunes. Dans ce document, nous allons décrire quelques-unes des faiblesses de la méthode telle qu'elle est appliquée au recensement. Dans la section 2, nous examinons les composantes de la CVD qui peuvent entraîner des erreurs dans les résultats finals. Dans la section 3, nous étudions les résultats de certaines analyses des estimations produites au moyen de la CVD ainsi que les données d'autres sources qui ont mis en évidence certains problèmes précis de la méthode aux fins de l'estimation de la population. La conclusion est présentée à la section 4.

2. FAIBLESSES DE LA CVD

Dans cette étude, on entend par faiblesse ou lacune tout ce qui restreint l'applicabilité des estimations établies au moyen de la CVD ou la confiance avec laquelle on peut les utiliser. Ces

Evaluation des estimations du sous-dénombrement obtenus par la contre-vérification des dossiers du recensement du Canada

R.D. BURGESS¹

RÉSUMÉ

Depuis 1961, Statistique Canada produit des estimations du sous-dénombrement pour chaque recensement. À cette fin, le Bureau utilise la méthode dite de contre-vérification des dossiers (CVD). La fiabilité des estimations est importante parce que ces dernières sont utilisées pour évaluer la qualité des données du recensement et pour établir les principales causes de l'erreur de couverture. Cette fiabilité a également un rôle très important dans l'élaboration de méthodes conçues pour améliorer la couverture des recensements futurs. Dans cette étude, nous définissons les sources potentielles d'erreur de la CVD de manière à bien les comprendre et, dans la mesure du possible, à en réduire les effets sur l'estimation de l'erreur de couverture.

MOTS CLÉS : Appariement; mobilité; biais de non-réponse; erreur de réponse; contre-vérification des dossiers; erreur d'échantillonnage; dépistage.

1. INTRODUCTION

Le recensement du Canada a lieu tous les cinq ans. Le dernier date de 1986. Depuis 1971, la principale méthode de dénombrement utilisée est l'autodénombrement; moins de 4% de la population est recensée par interview. Dans les régions où est appliquée la méthode de l'autodénombrement, les ménages sont inscrits sur une liste et un recenseur passe déposer un questionnaire chez ces derniers peu avant le jour du recensement (le 3 juin en 1976 et en 1981). Dans les grandes régions urbaines, on demande aux ménages de retourner leur questionnaire rempli par la poste au responsable local du dénombrement. Dans les régions rurales et les petites régions urbaines, un recenseur se rend chez les recensés pour reprendre les questionnaires. Le recenseur doit faire un certain nombre de vérifications de base en ce qui concerne la couverture et la qualité des réponses fournies par les ménages compris dans sa tâche. Les surveillants effectuent eux aussi certaines vérifications sur les questionnaires que leur remettent les recenseurs. Toutefois, la liste des ménages ne fait pas l'objet d'une vérification indépendante exhaustive. En outre, on a rarement la possibilité de vérifier le nombre de personnes inscrites dans le questionnaire de chaque ménage.

Il n'est donc pas étonnant qu'il y ait des erreurs de surdénombrement et de sous-dénombrement dans les chiffres du recensement. Ces erreurs sont importantes compte tenu des différentes utilisations des données du recensement. Par exemple, le nombre de députés pouvant être élus au Parlement du Canada est établi à partir des chiffres de population du recensement. De plus, pour calculer la participation financière des administrations fédérale et provinciales à différents programmes conjoints, on se fonde sur les chiffres de population ou la répartition de la population établis à partir du recensement (Statistique Canada 1983b). Enfin, la qualité des estimations de l'erreur de couverture est très importante du point de vue de l'utilité des données du recensement; de l'ajustement des chiffres de population et des logements pour tenir compte de l'erreur de couverture; et de l'amélioration possible de la

¹ R.D. Burgess, Division des méthodes d'enquêtes sociales, Statistique Canada, 4-ième étage, Immeuble Jean Talon, Tunney's Pasture, Ottawa, Ontario, K1A 0T6.

Des systèmes de codage automatique élaborés par des organismes centraux de statistique sont décrits dans deux articles. Lorigny traite du système QUID utilisé à l'Institut National de la Statistique et des Etudes Economiques. L'article de Wenzowski constitue un guide du système ACTR élaboré à Statistique Canada. QUID et ACTR sont tous deux conçus pour traiter de façon efficace tout genre de système de classification. Les lecteurs seront intéressés à comparer les méthodes utilisées par les deux systèmes. Certaines données sur le rendement sont aussi fournies. Mudryk décrit un système informatique pour le contrôle de la qualité actuellement en usage dans le cadre du programme global d'assurance de la qualité à Statistique Canada. Les objectifs du système visent à la fois à prévenir les erreurs dans les opérations de traitement des données d'enquête et à réduire progressivement les niveaux d'inspection à mesure que la qualité du traitement s'améliore et se stabilise.

Deguire décrit un système conçu pour analyser la syntaxe des adresses postales; ce système est en voie d'élaboration à Statistique Canada. Le logiciel produit des clés de recherche d'adresses formées de composantes d'adresses normalisées. Ces clés peuvent être employées au cours d'opérations d'appariement informatisé comme celles qui doivent être utilisées au cours de la construction d'un registre national d'adresses.

Emery décrit le SQL (Structured Query Language), le plus populaire de tous les langages d'interrogation utilisés avec les systèmes de gestion de bases de données relationnelles. Les points forts et les points faibles du langage sont soulignés.

Dans le dernier article de ce numéro, Nathan donne une liste de plus de 250 ouvrages, thèses et articles traitant de la méthode des réponses randomisées. L'article comprend aussi une classification par sujet.

Le rédacteur en chef

Dans ce numéro

Huit articles de ce numéro portent sur l'erreur de couverture dans le recensement. Ces articles, ainsi que les quatre qui examinent le même sujet et qui ont paru dans le numéro de juin 1988, fournissent au lecteur un bon aperçu de certaines des méthodes les plus récentes que l'on peut employer pour traiter cette question. Elle a dernièrement fait l'objet de beaucoup d'intérêt de la part des statisticiens et des décideurs. Dans de nombreux pays, des études sont faites pour estimer l'erreur de couverture soit pendant le recensement, soit après. Au Canada, la contre-vérification des dossiers (CVD) est la plus importante étude réalisée pour mesurer le sous-dénombrement. Aux États-Unis et en Australie, c'est une enquête postcensitaire (EP) qui est effectuée dans le même but.

Les articles de Burgess et de Romanuc portent sur les problèmes de couverture relatifs au recensement de la population du Canada. Burgess décrit la méthode de la CVD et il examine certaines des limitations de la CVD qui amènent des erreurs dans les estimations du sous-dénombrement. Romanuc, pour sa part, étudie la précision du recensement dans une optique démographique. Les résultats ainsi obtenus sont comparés à ceux qui proviennent de la CVD. De plus, Romanuc considère la qualité des composantes du changement de la population (naissances, décès, migration) utilisées dans la méthode démographique.

L'article de Choi, Steel et Skinner porte sur l'enquête postcensitaire (EP) effectuée en Australie en 1986. Comme Romanuc, les auteurs examinent les estimations du sous-dénombrement fondées sur l'analyse démographique. À la suite de leur analyse, les auteurs concluent qu'il faudra continuer d'avoir recours aux corrections basées sur l'EP pour le recensement de 1991, mais ils insistent sur le fait qu'il faut poursuivre les études portant sur les problèmes de biais. Cressie se sert d'un modèle pour estimer les erreurs dans le sous-dénombrement afin d'étudier les corrections à apporter aux chiffres du recensement. Il examine l'estimation synthétique, la méthode de Bayes et la méthode empirique de Bayes et il utilise la notion de risque pour comparer les estimateurs. On trouve que le risque le plus faible est associé à un estimateur empirique de Bayes ordinaire. Cressie fait remarquer que les résultats sont basés sur l'hypothèse qu'un nombre suffisamment élevé de ménages sont choisis pour l'EP.

L'article de Rubin, Schaffer et Schenker sur les méthodes d'imputation de valeurs manquantes dans l'enquête postcensitaire a aussi un certain contenu bayésien. Les auteurs examinent et critiquent les méthodes d'imputation étudiées par Schenker dans le numéro précédent de *Techniques d'enquête*. Ils proposent deux méthodes fondées sur les modèles et concluent que la méthode qui tient compte du mécanisme de non-réponse est préférable. Les auteurs préviennent les lecteurs que, même si leur méthode semble prometteuse, elle n'est pas encore tout à fait au point.

Fein et West examinent une classification systématique des causes du sous-dénombrement et ils concluent que l'oubli de personnes dans des ménages est la cause majeure du sous-dénombrement. Mulry et Spencer présentent une analyse méthodologique de l'erreur totale de l'estimateur de système dual (un estimateur traité par des auteurs dans le numéro de juin 1988). Les auteurs utilisent une méthode bayésienne pour combiner les composantes d'erreur afin d'obtenir une estimation d'intervalle final du taux de sous-dénombrement net. Zaslavsky traite du problème du sous-dénombrement à l'aide d'estimations du sous-dénombrement pour les ilots afin de répondre les ménages de l'ilot. Cette méthode offre l'avantage de préserver le "caractère" de chaque ilot. Les détails de la méthode sont intéressants et sembleront familiers aux lecteurs qui connaissent la méthode itérative du quotient.

L'élaboration de nouveaux systèmes informatiques conçus pour traiter de grandes quantités de renseignements constitue un sujet qui présente un intérêt croissant pour les statisticiens qui s'occupent d'enquêtes. Cinq des articles de ce numéro décrivent des logiciels élaborés en fonction des techniques d'enquête.

TABLE DES MATIÈRES – suite

Développements de logiciels

J. LORIGNY	307
QUID, une méthode générale de chiffrement automatique	307

M.J. WENZOWSKI	317
ACTR: Un système généralisé de codage automatique	317

W. MUDRYK	327
Système de gestion de la qualité dans les opérations d'enquêtes	327

Y. DeGUIRE	335
Analyse des adresses postales	335

D.N. EMERY	345
Note d'information sur SQL	345

G. NATHAN	351
Bibliographie de la méthode des réponses randomisées: 1965-1987	351

Rectification	367
---------------	-----

Remerciements	369
---------------	-----

TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada
Volume 14, numéro 2, décembre 1988

TABLE DES MATIÈRES

Dans ce numéro	145
Section Spéciale - Erreur de couverture dans le recensement	

R.D. BURGESS	Évaluation des estimations du sous-dénombrement obtenus par la contre- vérification des dossiers du recensement du Canada	147
--------------	------------------------------------------------------------------------------------------------------------------------------------	-----

A. ROMANUC	Une approche démographique à l'évaluation du recensement de 1986 et des estimations de population pour le Canada	169
------------	---------------------------------------------------------------------------------------------------------------------------	-----

C.Y. CHOI, D.G. STEEL, et T.J. SKINNER	Redressement des chiffres du recensement de 1986 en Australie pour le sous-dénombrement	187
----------------------------------------	--------------------------------------------------------------------------------------------------	-----

N. CRESSIE	Dans quelles circonstances les opérations de redressement améliorent-elles les chiffres du recensement?	205
------------	------------------------------------------------------------------------------------------------------------------	-----

D.B. RUBIN, J.L. SCHAFER, et N. SCHENKER	Méthodes d'imputation de valeurs manquantes dans des enquêtes postcensitaires ..	223
------------------------------------------	----------------------------------------------------------------------------------	-----

D.J. FEIN et K.K. WEST	Sources du sous-dénombrement lors du recensement: Résultats du recensement d'essai de 1986 à Los Angeles	237
------------------------	-------------------------------------------------------------------------------------------------------------------	-----

M.H. MULRY et B.D. SPENCER	L'erreur totale dans l'estimateur de système dual: Recensement du Central Los Angeles County de 1986	257
----------------------------	---------------------------------------------------------------------------------------------------------------	-----

A.M. ZASLAVSKY	Redressement des estimations régionales par une repondération des ménages.....	281
----------------	--------------------------------------------------------------------------------	-----

TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada

La revue Techniques d'enquête est répertoriée dans The Survey Statistician et Statistical Theory and Methods Abstracts. On peut en trouver les références dans Current Index to Statistics.

COMITÉ DE DIRECTION

Président

G.J. Brackstone

Membres

B.N. Chinnappa

G.J.C. Hole

C. Patrick

F. Mayda (Directeur de la production)

COMITÉ DE RÉDACTION

Rédacteur en chef

M.P. Singh, *Statistique Canada*

Rédacteurs associés

K.G. Basavarajappa, *Statistique Canada*

D.R. Bellhouse, *U. of Western Ontario*

L. Biggert, *Université de Florence*

D. Binder, *Statistique Canada*

E.B. Dagum, *Statistique Canada*

W.A. Fuller, *Iowa State University*

J.F. Gentleman, *Statistique Canada*

M. Gonzalez, *U.S. Office of Management and Budget*

D. Holt, *University of Southampton*

G. Kalton, *University of Michigan*
M.N. Murthy, *Applied Statistics Centre, India*
W.M. Podehl, *Statistique Canada*
J.N.K. Rao, *Carleton University*
D.B. Rubin, *Harvard University*
I. Sande, *Statistique Canada*
C.E. Sarnadal, *Université de Montréal*
F.J. Scheuren, *U.S. Internal Revenue Service*
V. Tremblay, *Statplus, Montréal*
K.M. Wolter, *U.S. Bureau of the Census*

POLITIQUE DE RÉDACTION

La revue Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contreparties d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

La revue Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à en faire parvenir le texte au rédacteur en chef, M. M.P. Singh, Division des méthodes d'enquêtes sociales, Statistique Canada, 4^e étage, Edifice Jean-Talon, Tunney's Pasture, Ottawa (Ontario), Canada KIA 0T6. Prière d'envoyer deux exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

Abonnement

Le prix de la revue Techniques d'enquête (catalogue n° 12-001) est de 30,00\$ par année au Canada, et de 35,00\$ par année à l'étranger (paiement en dollars canadiens ou l'équivalent). Prière de faire parvenir votre demande d'abonnement à: Section des ventes des publications, Statistique Canada, Ottawa (Ontario), Canada KIA 0T6. Un prix réduit, soit 16,00\$ (E.-U.) (20,00\$ Can.) est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête et la Société Statistique du Canada. Veuillez envoyer votre demande d'abonnement directement à l'organisation.

TECHNIQUES D'ENQUÊTE

UNE REVUE DE STATISTIQUE CANADA

DÉCEMBRE 1988

Publication autorisée par
le ministre de l'Expansion industrielle régionale et
le ministre d'État, sciences et technologies

©Ministre des Approvisionnements
et Services Canada 1988

Le lecteur peut reproduire sans autorisation des
extraits de cette publication à des fins d'utilisation
personnelle à condition d'indiquer la source en
entier. Toutefois, la reproduction de cette publication
en tout ou en partie à des fins commerciales ou de
redistribution nécessite l'obtention au préalable d'une
autorisation écrite du Groupe des programmes et produits
d'édition, agent intermédiaire aux permis, administration
des droits d'auteur de la Couronne, Centre d'édition
du gouvernement du Canada, Ottawa, Canada K1A 0S9.

Mars 1989

Prix: Canada, \$30.00 par année
Autres pays, \$35.00 par année

Paiement en dollars canadiens ou l'équivalent

Catalogue 12-001, vol. 14, n° 2

ISSN 0714-0045

Ottawa

Canada

VOLUME 14, NUMÉRO 2
DÉCEMBRE 1988

UNE REVUE
DE
STATISTIQUE CANADA

TECHNIQUES D'ENQUÊTE



JUL 19 1989

